ARL-TR-8027 • May 2017

US Army Research Laboratory

Heterogeneous Systems for Information-Variable Environments (HIVE)

by Amar R Marathe, Benjamin T Files, Jonroy D Canady, Kim A Drnec, Hyungtae Lee, Heesung Kwon, Allison Mathis, William D Nothwang, Garrett Warnell, and Ethan Stump

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-8027 ● May 2017

US Army Research Laboratory

Heterogeneous Systems for Information-Variable Environments (HIVE)

by Amar R Marathe, Benjamin T Files, Jonroy D Canady, and Kim A Drnec
Human Research and Engineering Directorate, ARL

Hyungtae Lee, Heesung Kwon, Allison Mathis, and William D Nothwang
Sensors and Electron Devices Directorate, ARL

Garrett Warnell and Ethan Stump
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
May 2017	DSI Report	October 2013–September 2016

4. TITLE AND SUBTITLE
Heterogeneous Systems for Information-Variable Environments (HIVE)

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Amar R Marathe, Benjamin T Files, Jonroy D Canady, Kim A Drnec, Hyungtae Lee, Heesung Kwon, Allison Mathis, William D Nothwang, Garrett Warnell, and Ethan Stump

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Research Laboratory
ATTN: RDRL-HRF-D
Aberdeen Proving Ground, MD 21005-5425

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-8027

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The US Department of Defense envisions the use of autonomy (independent robotic systems) in complex operational environments where there are substantial constraints on communications, sensing capability, and local processing. Many instances of autonomous systems exist that function very well in highly constrained situations and environments but fail to generalize to novel unconstrained environments. As such, foreseeable future autonomy will not have the ability to independently function in these scenarios. This report describes fundamental research into the issues underlying the control of such heterogeneous systems and aims to lay the groundwork for the development of generalizable methodologies for the effective integration of heterogeneous agents in dynamic, information-variable environments. Specifically, we focus on 4 research areas: 1) directly accounting for human variability to enable better integration of human decisions, 2) data fusion/computer vision (CV), 3) agent adaptation, 4) and networking dynamic teams of humans and machines. Additionally, we describe efforts to integrate these 4 research areas through both ongoing experiments and plans for future work. In the long term, these theories will enable the creation of an analytical framework for human–machine network design and control and enable powerful, highly adaptive human–autonomy systems that will be a major technological push affecting a broad range of applications.

15. SUBJECT TERMS
Director's Strategic Initiative (DSI), human–autonomy integration, human variability, CV, navigation, distributed teaming, intelligent agent adaption

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Amar R Marathe
Unclassified	Unclassified	Unclassified	UU	180	19b. TELEPHONE NUMBER (Include area code) (410) 278-3638

Contents

List of Figures

List of Tables

INTENTIONALLY LEFT BLANK.

1. Introduction

The US Department of Defense (DOD) envisions the use of autonomy (independent robotic systems) in dynamic and complex operational environments where there are substantial constraints on communications, sensing capability, and local processing, and many instances of autonomous systems exist that function very well in highly constrained situations and environments. Foreseeable future autonomy still will not have the ability to independently function in these scenarios. Therefore, to bridge this gap, we will need to rely on heterogeneous human–autonomy systems for the effective integration of autonomy (e.g., the highly networked Soldier Future Squad with its accompanying few small robotic systems). This report describes fundamental research into the issues underlying the control of such heterogeneous systems and aims to lay the groundwork for the development of generalizable methodologies for the effective integration of heterogeneous agents in dynamic, information-variable environments.

As various autonomies are proliferating across numerous DOD missions, human–autonomy integration is still largely implemented through hierarchical schemes that maintain the human as the ultimate decision authority. These frameworks in which humans must constantly monitor autonomous systems inadequately leverage the unique capabilities of different agents. Further, the substantial constraints on communications and the well-documented overloading of the Soldier are prohibitive to the employment of these approaches for even small heterogeneous teams. We hypothesized that 3 deficiencies on current control architectures have limited the ability of heterogeneous human–autonomy systems to adequately share decision-making authority between humans and machines. Specifically, control architectures have not been designed to 1) account for performance variability across and within human agents, 2) foster positive joint adaptation by human and autonomy to situation dynamics, and 3) enable robust human–machine interaction under changing communications, sensor, and processing constraints.

Over the past 2 decades, there have been substantial research advancements that may lead to methodologies to address the deficiency in accounting for performance variability across and within human agents. This research has shown that humans are widely variable both between subjects and even within themselves due to an array of factors. This variability, which emerges across a wide range of time scales, is often ignored in human–autonomy systems.[1] Current technologies to estimate individual humans' abilities are providing more-precise capabilities that potentially could be incorporated into control architectures.[2] However, even with near-future

technologies, such estimates will still include substantial signal uncertainty. Potential approaches to overcoming this issue may be found in the sensor fusion and human–autonomy-teaming research communities, which have illuminated approaches to joint estimation for combined decision making. These approaches show how multiple, disparate sensor systems with sometimes substantial uncertainty are integrated to yield reliable state estimates.[3–5] However, these concepts have only focused on uncertainty in autonomy and have neither been extended across broad time scales nor have they been applied to human–autonomy systems.

Regarding the deficiency in fostering positive joint adaptation by human and autonomy to situation dynamics, beyond the frequent use of modified hierarchical control schemes, human–autonomy estimations have largely not been addressed in the literature.[4,6] This is due in part to a lack of generally accepted models for human decision making.[7,8] It is well understood, though, that human decisions can be biased depending on the information they are given a priori.[9] Robotic systems, conversely, often share a common decision-making model[10] or at least an understanding of decision making in adjacent nodes,[11] and, consequently, there has been substantial work on distributed decision making for autonomous systems under varying constraints.[12] These distributed decision-making methods for autonomous systems often suffer from reinforcement of bad data[13] and do not deal well with highly dynamic environments when there is communications latency.[14]

Regarding the deficiency in robust human–machine interaction under changing constraints, an important question in system integration is how to merge the outputs of many independent, networked agents into a cohesive output. While weighted sum models, such as Kalman Filters[15] and Bayesian approaches,[16] are widely used for intranode decision making, they have been ineffective when applied to distributed systems. In the realm of autonomy, the 2 major approaches to addressing this issue are to either solve a distributed optimization problem to explicitly synchronize several agents, usually through an efficient mechanism such as auctions,[17] or to let agents reach a stable solution through local interactions, as in the biologically inspired “flocking” techniques.[18]

2. Approach

We seek to enable the future of networked heterogeneous human–autonomy systems by addressing the 3 aforementioned deficiencies in current control architectures. In directly addressing the deficiencies, our approach has the following 4 broad research areas:

1) *Directly accounting for human variability to enable better integration of human decisions*. Human decision-making performance is highly variable and human decisions are fallible. Over the long term, we aim to establish probabilistic approaches to estimate human decision-making performance across a broad range of time scales. Such measures allow us to capture both variation among individuals and an assessment of the changing capabilities within an individual through time. We focused on metrics at multiple time scales and sought to validate both measures and confidence scores estimating human decision-making performance at each of these time scales.

2) *Data Fusion/Computer Vision.* A critical barrier for fielding autonomous systems is the issue of human–autonomy integration. Effective methods for fusing information across human and autonomous agents are not yet robust to dynamics in human performance and in the environment. This effort focused on computer vision (CV) as the exemplar autonomous technology. Recently CV algorithms have dramatically improved, enabling an unprecedented level of accuracy in understanding the contents of images. Nevertheless, most algorithms are unable to function in highly dynamic and cluttered environments. We developed novel fusion approaches to extract complementary information from multiple types of CV algorithms and leveraged these fusion architectures to also integrate CV and human inputs. This approach can be leveraged to incorporate information about the dynamics in both the environment and human performance in an effort to rapidly adapt to these changes in real time.

3) *Agent Adaptation*. In adaptive systems, dynamics in both the environment and in human performance can lead to a breakdown in performance for autonomous systems. Ultimately, we aim to develop novel approaches to enable the autonomous agents to dynamically adapt to both human and environmental variability. In this project, we focused on adaptations to dynamics in the environment.

4) *Networking dynamic teams of humans and machines*. Building on objectives 1 and 2, we seek to develop control and networking policies that best facilitate interactions within human–autonomy systems. These must adapt to, and be robust to sensor noise, scenario dynamics, competing priorities, intermittent communications, and processing limitations. Ultimately, we aim to develop systems that share authority among networked, heterogeneous human and autonomous agents and create an analytical framework for networked human–machine decision making and control. In this project, we focused on control and networking policies for a small defined network functioning in specific contexts.

3. Human Variability

Human decisions and actions are highly variable and this variability encompasses performance ranging from exceptionally good to exceptionally bad. This variability can complicate efforts to develop collaborative methods for human–autonomy interaction (HAI) by reducing the predictability of human performance. For example, human psychological and physiological states have been observed to vary considerably within an individual operator, even over short time scales[1] or across immediately successive actions.[19] In order to integrate human inputs with autonomy to enhance performance in complex, dynamic circumstances, system designers must develop strategies to account for and exploit this variability. To do so would enable powerful adaptive systems that leverage the unique strengths of each agent, while offsetting instances where their decisions or actions lead to potential failure or catastrophe.

There has been considerable research into mitigating the potential impact of human variability and performance failures on HAI systems. The extant literature has most commonly offered substitution-based function allocation to toggle exclusive control or decision authority between humans and autonomous systems.[20,21] Some function allocation concepts have considered task type and the level of autonomy[22] alongside typical "man-is-better-at"–"machine-is-better-at" roles.[23] Such function-allocation concepts have been instantiated in a number of different control frameworks, the most widely recognized of which is supervisory control.[24] The supervisory control framework can be implemented in a variety of ways, ranging from autonomous waypoint navigation to shared control schemes in which both the human and the autonomous system provide control inputs with different relative contributions (e.g., Crandall and Goodrich[25]). Adaptive schemes have also been developed to enable active management of the balance of inputs from human and autonomous agents through user selection,[26] based on cost-benefit estimates of the performance of the agents,[27] or by enabling the autonomy to periodically query the operator for assistance.[28,29] Unfortunately, the majority of these approaches have only succeeded in limited and controlled contexts and have not been widely adopted for real-world use.

To overcome the issue of unpredictable human performance for HAI systems, we developed methods to characterize and quantify the reliability of human performance. Reliability was operationalized as measures of confidence. Decisions made during periods of low-reliability performance by one agent produced a lower confidence, which reduced the impact of those decisions on the overall joint decision. This is not a novel concept; rather, this is a fundamental aspect of decision theory.[30–32] The goal of developing measures of confidence in human performance

was inspired by prior work on sensor fusion and human–autonomy integration that has shown how multiple, disparate sensor systems with sometimes substantial uncertainty could be integrated to yield reliable decisions.[4,33]

Confidence metrics are typically derived from statistical uncertainty measures and can be directly integrated into a control system. The novel aspect here was in the application of confidence to human inputs as well as to data from other sensors. Several challenges must be met to develop and validate appropriate confidence metrics for human data. Given that human sources of input are typically treated as having little or no noise, they have historically been intentionally constrained to a level, such as a button press (BP), that was presumed to be unambiguous and effectively without noise. However obtained, human inputs have most often been trusted and then integrated as they were received.[28,29,34,35] While it is well understood that human psychological and physiological states vary widely, both across and within individuals,[1,19] it remains less well understood how to predict the expected variability given an observed state in a specific person working within a particular task environment. This is due in large part to an incomplete understanding of how states observed in similar contexts will change over time as well as across and within individuals.

These challenges are further exacerbated by current human sensing techniques that produce data that are inconsistent, invalid, or both, when applied in real-world circumstances.[36] As such, the available measures based on human-sensed data, including those from overt behavior and physiology, have not yet been widely integrated into human–autonomy systems.[1,37]

In this effort, we focused our work on human variability using a 3-step process. First, we sought to characterize human performance on multiple time scales. Next, on the basis of this characterization, we identified behavioral and physiological features that relate to changes in task performance. Finally, we used these features to develop confidence measures that predict the reliability of task-related human performance. We focused this effort on 2 primary testbed tasks: target identification and exploration.

Testbed Task 1: Rapid Serial Visual Presentation (RSVP)/Target ID

Background

Finding target images in large databases of candidate images is a difficult problem, and while CV algorithms are adequate for some tasks, for others human vision is required. A key insight to approaching this problem is that humans tasked with finding target images achieve high target-detection accuracy even if the images are shown very rapidly.[38] Using RSVP with images displayed at rates of 2–10 Hz can

dramatically increase the rate at which target images are found in image databases compared to self-paced image viewing.[39,40] To optimally incorporate human image processing into a heterogeneous team, a confidence metric is needed for human performance on RSVP target-identification tasks. We took 2 parallel approaches to developing confidence measures. First, we developed a method for improved quantification of human performance in a standard RSVP task.[41] Second, we examined RSVP performance on a task with increased complexity, and will show how the inclusion of rudimentary confidence measures can improve performance.[42]

3.1 Improved Characterization of Human Performance in RSVP

RSVP target-detection task performance can be difficult to quantify due to response-time variability.[39,43] Here we introduce a novel method for estimating performance on the RSVP target-detection task in settings in which image labels are known. This improved performance estimate can be derived from a training set of images with known labels.

RSVP target-detection performance can be quantified by the subject's hit rate (HR) and false alarm rate (FAR). Knowing whether a response is a hit or a false alarm requires knowing whether a target or a nontarget stimulus evoked the response. Because of response-time variability, it can be difficult to know what stimulus evoked a button-press response. For example, a response might be a relatively fast response to a target stimulus or a relatively slow response to the preceding nontarget stimulus. When the response-time variability substantially exceeds the inter-stimulus interval, situations arise in which a response could just as easily be attributed to any of several stimuli. One method currently in use for estimating HR and FAR entails establishing a temporal window after each target stimulus (e.g., 0–1 s relative to target onset) and declaring any response that falls in that window is a hit. Other methods estimate a response-time probability density function and use that to assign responses to stimuli.

Under this program we developed a method of performance estimation that generally outperforms other methods currently in use for estimating the HR and FAR in RSVP target-detection tasks. Using simulations with known HRs and FARs, we showed that our method is more accurate than established methods. This advantage is especially clear when the stimulus presentation rate is high and/or the FAR is nonzero.

3.1.1 Established Methods for Estimating HR and FAR

There are 2 classes of methods for determining HR and FAR in common use with RSVP target-detection tasks. The first uses a windowing approach, establishing a minimum and a maximum response time, typically from 0 to 1000 ms posttarget. Any response that falls within that window after a target is declared a hit, and then the HR is determined as the number of declared hits divided by the total number of targets. Responses that do not fall within a window corresponding to any target are declared false alarms, and the FAR is calculated as the number of false alarms divided by the number of nontarget stimuli (Fig. 1). Implementations of this method differ in how responses are scored when more than one response falls within a response window and/or what to do when a response falls within more than one response window.

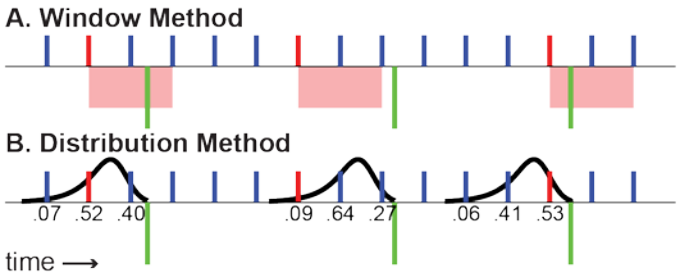

Fig. 1 Timelines illustrating existing response-assignment methods: blue hash marks indicate onset times of nontarget stimuli; red hash marks indicate onset times of target stimuli; downward green hashes indicate times a response occurred. Interstimulus interval is 0.5 s. In the A) window method, a window of time (typically 0–1 s posttarget) is established; responses falling within that window are declared hits. Here, the first and third responses would be classified as hits, second would be a false alarm. Same experimental timeline as analyzed with the B) distribution method: black curves are the response-time probability density function reversed and with its origin at the times of response; numbers below stimulus hashes show attribution resulting from the corresponding response.* Using maximum likelihood (max method) assigns response to the stimulus with the highest likelihood.

The second class of methods for estimating HR and FAR uses a response-time distribution to estimate a response-time probability density function (RT-PDF) that is used to assign responses to specific stimuli.[44] The likelihood that a BP was in response to a specific candidate stimulus is estimated as the probability of that particular response time relative to the time of the candidate stimulus (i.e., the estimated value of the RT-PDF). The likelihood is then normalized by dividing the likelihood for each candidate stimulus by the sum of the likelihoods for all candidate stimuli.[45] From here the methods in this class diverge. One approach is to assign responsibility for the response to the stimulus with the maximum

* Corresponding response as computed using Eq. 1 on page 8.

likelihood. If that stimulus is a target, the response is counted as a hit; if the stimulus is a nontarget, the response is counted as a false alarm. The other approach is to distribute responsibility for the response to various stimuli according to the normalized likelihood that they generated the response. Because the distribution method is central to the method proposed in this report, it will be useful to define the function used to distribute responsibility, called here the *apportionment* function. Given times of stimulation *S*, a stimulus of interest at time S_i, a response at time *T*, and an RT-PDF function *f*, the apportionment function is defined as

$$A(S_i, T) = \frac{f(T-S_i)}{\sum_j f(T-S_j)}. \quad (1)$$

Using this approach, if the apportionment worked out such that 0.52 of the response was apportioned to a target stimulus and the remaining 0.48 was apportioned to a nontarget stimulus, that response would count as 0.52 of a hit and 0.48 of a false alarm (as illustrated in Fig. 1).

3.1.2 Proposed Method

The regression method introduced here is based on the aforementioned apportionment method (Eq. 1). The proposed method estimates the expected response apportionment to each stimulus as a function of the probability that nearby stimuli will generate responses and the proportion of those possible responses that will be apportioned to the stimulus of interest. The expected response apportionment for the *i*th stimulus is the sum of the expected apportionment due to responses to all nearby stimuli, S_j.

$$E[A(S_i)] = \sum_j E[A_s(S_j, S_i)], \quad (2)$$

where $A_s(S_j,S_i)$ is similar to $A(S_i)$ but only computes the attribution onto S_i of responses actually generated by S_j. The expected value of $A_s(S_j,S_i)$ is

$$E\left[A_s(S_j, S_i)\right] = \sum_T p(T)A(S_i, T). \quad (3)$$

The term $p(T)$ is the probability that a response elicited by $\underline{S_j}$ occurs at time *T*. This term can be split into the probability that any response is elicited by stimulus S_j, denoted $p(R|S_j)$, times the probability that a response occurs at a specific time. The latter quantity is obtained from the response-time probability density function, *f*.

$$E\left[A_s(S_j, S_i)\right] = p(R|S_j) \sum_T \left[f(S_j - T)A(S_i, T)\right]. \quad (4)$$

Substituting this equation into Eq. 2 yields the following:

$$E[A(S_i)] = \sum_j \left(p\left(R\middle|S_j\right) \sum_T \left[f\left(S_j - T\right)A(S_i, T)\right]\right). \quad (5)$$

Note that for simplicity of notation the limits of summation for *j* and *T* are not given. However, *f(x)* is zero for negative *x* and approaches zero as *x* increases, and $A(S_i,T)$ goes to zero as S_i-*T* increases in magnitude, so in practice, only a limited range of *j* and *T* need to be calculated.

This equation can be simplified under the assumptions of a typical RSVP target-detection experiment; namely, there are stimuli that are targets and stimuli that are nontargets, and that the probability of responding to a target is a constant hit rate *HR*, and the probability of responding to a nontarget is a constant false alarm rate *FAR*. If the stimulus at S_j is a target, then $p(R|S_j, S_j \in tar)$ is *HR*. If the stimulus at S_j is a nontarget, then $p(R|S_j, S_j \in n.t.)$ is *FAR*. Separating out the target and nontarget stimuli near the stimulus of interest, the equation becomes the following:

$$E[A(S_i)] = HR \times \sum_{S_j \in tar} \sum_T \left[f\left(S_j - T\right)A(S_i, T)\right] + FAR \times \sum_{S_j \in n.t.} \sum_T \left[f\left(S_j - T\right)A(S_i, T)\right]. \quad (6)$$

For each stimulus in the experiment, both summation terms can be computed based on the known stimulus timings and the RT-PDF. This yields a system of simple linear equations with one equation per stimulus with 2 unknowns: HR and FAR. Least-squares linear regression can then be used to find the values of HR and FAR that best fit the observed attribution for each stimulus; these are the estimates of the HR and FAR for the experiment.

3.1.3 Evaluation Methods

Having introduced the mechanics of the proposed method, simulations are described that compare the performance of the proposed method with state-of-the-art methods. The general approach was to simulate responses based on a known HR and FAR and then analyze the simulated data using the proposed method as well as the 3 other methods for estimating HR and FAR previously described (Fig. 2). To ensure that the stimulation timeline we used was wellfounded, we used the timeline of stimulus and response events from a RSVP target-detection experiment that has been described previously.[46–48] Portions of the methods of that experiment are summarized here because the stimulus timeline and response-time distributions from that experiment were used in our simulations.

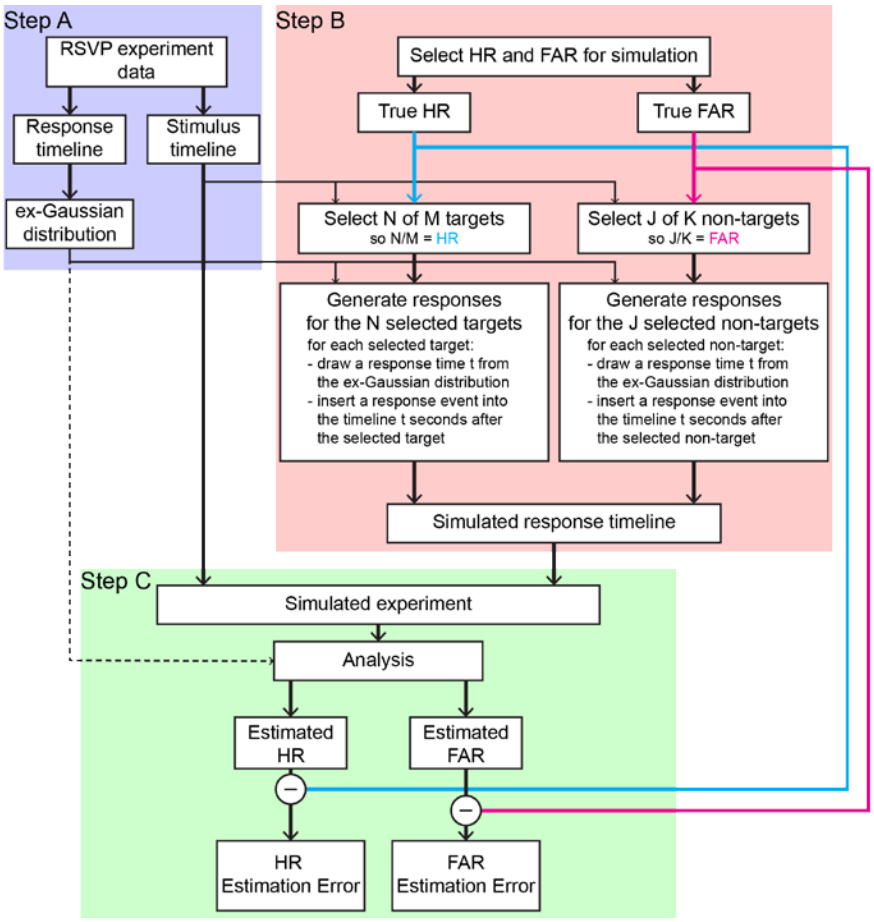

Fig. 2 Simulation method: the process for one iteration of the simulation. This process was repeated 250 times per combination of HR and FAR. Analysis was done separately using each of the 4 analytical methods described previously.

3.1.4 Participants

Fifteen participants (9 male, 6 female, ages 18–57, average 39.5) volunteered for the current study. Participants provided written informed consent, reported normal or corrected-to-normal vision, and reported no history of neurological problems. Fourteen of the 15 participants were right-handed. The voluntary, fully informed consent of the persons used in this research was obtained as required by federal and Army regulations.[49,50] The investigator has adhered to Army policies for the protection of human subjects.[50]

3.1.5 Stimuli and Procedure

Stimuli consisted of short video clips that contained either people or vehicles in background scenes (target stimuli), or only background scenes (nontarget stimuli). Participants were instructed to make a manual BP with their dominant hand immediately when they detected a target, and to abstain from responding to

nontarget stimuli. Video clips consisted of 5 consecutive images each 100 ms in duration; each video clip was presented for 500 ms. There was no interval between videos such that the first frame was presented immediately after the last frame of the prior video. If a target appeared in the video clip, it was present on each 100-ms image. The nontarget-to-target ratio was 90/10. RSVP sequences were presented in 2-min blocks, after which time participants were given a short break. Participants completed a total of 25 blocks.

3.1.6 Simulations

3.1.6.1 Extracting a Response Time Probability Density Function

All simulations and analyses were done using custom scripts in MATLAB version 2014a (MathWorks, Natick, Massachusetts).The RT-PDF used in the simulations was derived from the responses in the original timeline (Fig. 2, Step A). An empirical response-time distribution was created by iterating over all target stimuli and looking for any response that fell between 200 and 1500 ms after the target. The latency of responses relative to the associated target events was then fit with an ex-Gaussian distribution using maximum-likelihood estimation.[51] The ex-Gaussian distribution is the sum of an exponential and a Gaussian; this distribution was selected because it compactly describes empirical response-time distributions reasonably well.[52] After estimating the RT-PDF, the responses in the original timeline were no longer considered for the simulations.

3.1.6.2 Simulating Responses

Several simulations were then run to determine the accuracy with which the estimation methods described above recover the HR and FAR under different true values of those quantities. A total of 101 HRs, ranging uniformly from 0 to 1, were combined with 101 FARs, also ranging uniformly from 0 to 1, resulting in 10,201 combinations of HR and FAR. To collect statistics on the performance at each combination of HR and FAR, each simulation was repeated 250 times.

For each simulation, an HR and FAR were selected (Fig. 2, Step B). Then, a random subset of all targets and nontargets was selected to generate responses such that the simulated rates were as close as possible to the selected rates (while still having whole numbers of responses). When a response was generated, a random draw was taken from the response-time distribution (as described by the RT-PDF), and a response event was added at that time after the generating stimulus.

3.1.6.3 Analyzing the Simulated Experiment

After simulating all of the responses necessary to generate the target HRs and FARs, the stimulus and simulated response timeline were analyzed using the 4 methods previously described: the window method, the maximum likelihood method (max), the distribution method, and the regression method (Fig. 2, Step C). Three stimulus presentation rates (stimuli per second) were simulated as well: 2, 4, and 10 Hz. The original experiment used a presentation rate of 2 Hz. To simulate faster presentation rates, the sampling rate of the experiment was multiplied by 2 and 5, respectively, while leaving the response-time distribution unchanged. This guaranteed that any change in the HR and FAR estimates was due to the presentation rate and not a difference in the total number of stimuli.

Three of the 4 methods tested (all but the window method) make use of the RT-PDF. In the first round of simulations, these 3 methods used the same RT-PDF that generated the data. In an experimental setting, however, the RT-PDF is not known a priori and must be estimated. When the HR is high enough and the FAR is low enough, an RT-PDF can be estimated from the data, as outlined previously. However, if the HR is suspected to be low, the method may produce an inaccurate estimate of the RT-PDF. We wanted to examine the relative performance of these methods when the RT-PDF cannot be estimated. In the second round of simulations, to simulate a worst-case scenario, the 3 methods that rely on an RT-PDF estimate were provided an RT-PDF that was uniform over the interval [0, 1000 ms]. That interval was chosen to correspond to the interval used by the window method. This flat RT-PDF introduces a high probability of multiple stimuli receiving equal attribution for a given response. This is relevant to the max method, because it assigns full attribution to the stimulus with maximal attribution. To resolve ties, the max method attributes the response to the earliest stimulus with maximal attribution.

Then, to examine the impact that the choice of method for HR and FAR estimation can have on experimental results, the HR and FAR were estimated using the actual (rather than simulated) response data.

3.1.7 Results

For each simulation, the HR and FAR estimation errors were computed as the difference between the simulated rate and the rate estimated by the estimation method under examination. For example, if the true HR were 0.8 but the method estimated the HR to be 0.75, the estimation error would be –0.05.

The remainder of the results section is organized as follows: First, an illustrative subset of the simulation results is presented. This subset was chosen to show

simulation results for HRs and FARs that might be obtained with poor, good, or excellent target detection performance. Second, all of the simulation results are summarized to provide a comprehensive overview of the performance of these 4 estimation methods. Third, results of statistical tests are presented that tested for bias in the estimation methods used. Fourth, the results of simulations with an inaccurate RT-PDF are summarized. Finally, the results of applying each of the 4 estimation methods to real (rather than simulated) RSVP target-detection data are shown to illustrate the practical impact that the choice of estimation method can have.

3.1.7.1 An Illustrative Subset of Results

Although actual performance in RSVP target detection will depend heavily on the stimuli, task, and participant, 3 pairs of HR and FAR were chosen as illustrative exemplars of poor (HR 0.50, FAR 0.10), good (HR 0.80, FAR 0.02), and excellent (HR 0.99, FAR 0.01) performance (Fig. 3). Overall, when HR is high and FAR is low (i.e., in the good and excellent performances), the distribution and max methods make larger systematic errors than the other 2 methods, and the window method makes errors comparable to the regression method. As the presentation rate increases, the difference in the relative performance increases as well. In the poor performance case, the errors made by the regression method are clearly smaller than the others except at the 2-Hz presentation rate. At that rate, the regression, max, and distribution methods make comparable errors that are smaller than the errors made by the window method.

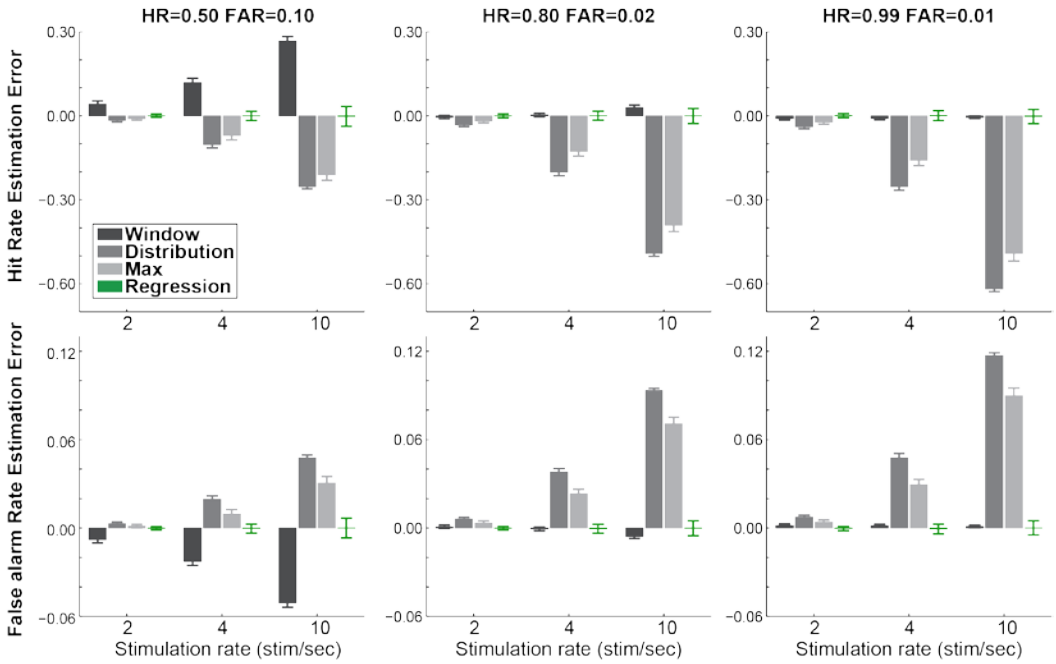

Fig. 3 Estimation method performance examples: plots illustrate estimation results for specific combinations of simulated HR and FAR when the true probability mass function of the response times was known. The pairs HR 0.50, FAR 0.10; HR 0.80, FAR 0.02; and HR 0.99, FAR 0.01 were selected as illustrative of poor, good, and excellent RSVP target-detection performance, respectively. Bars show the median estimation error, and error bars show ±1 standard deviation for each of the 4 estimation methods at 3 presentation rates. The upper row of panels shows HR estimation errors; the lower row shows FAR estimation errors.

3.1.7.2 Full results

Considering the full range of simulated HRs and FARs, for all but the regression method, substantial systematic errors were apparent that depended on a combination of the simulated HR, simulated FAR, and simulated presentation rate for HR estimation (Fig. 4) and FAR estimation (Fig. 5). Estimation errors taken over the entire range of simulated HR and FAR were smallest for the regression method at all simulated rates with median absolute difference among estimated and simulated HRs of 0.001, 0.002, and 0.004 for presentation rates of 2, 4, and 10 Hz, respectively (Table 1), and median absolute difference for FARs of 0.001, 0.002, and 0.002 for presentation rates of 2, 4, and 10 Hz, respectively (Table 2). However, the regression method also had the largest variability for HR estimates at 4 and 10 Hz presentation and FAR at 10 Hz, measured as the standard deviation of all estimates after the median of all 250 estimates within a simulated HR/FAR cell had been subtracted (Tables 1 and 2). For the HR estimate, the regression method's variability was 0.022 and 0.053 at 4 and 10 Hz, respectively, and for the FAR estimate, the regression method's variability was 0.010 at 10 Hz.

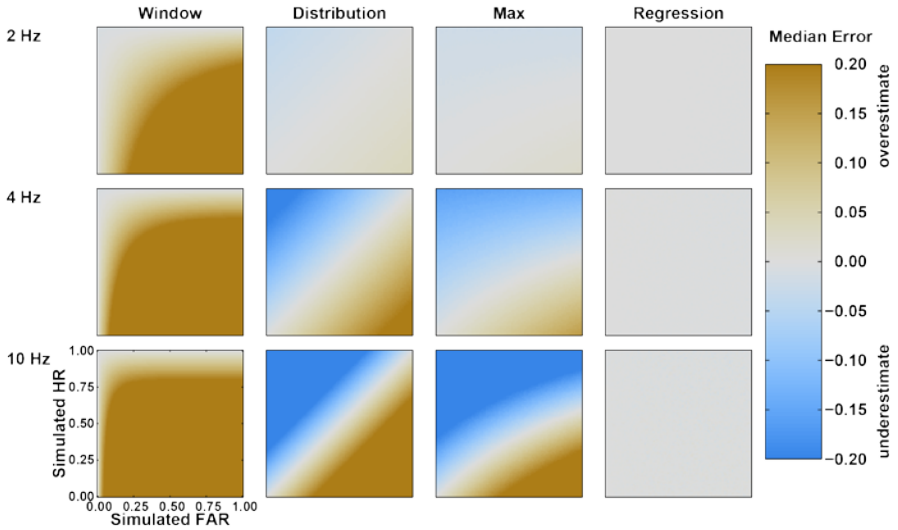

Fig. 4 HR estimation error summary: panels show simulation results for estimation methods (columns) at a particular presentation rate (rows) when the true probability density function of the response times was known. Colors indicate the difference between median estimate of the HR and simulated value of HR clipped to an absolute value of 0.2. Within a panel, simulated FAR increases from left to right; simulated HR increases from bottom to top. All methods except regression have HR estimation errors that clearly depend on the simulated HR and FAR, and overall magnitude of errors increases as presentation rate increases.

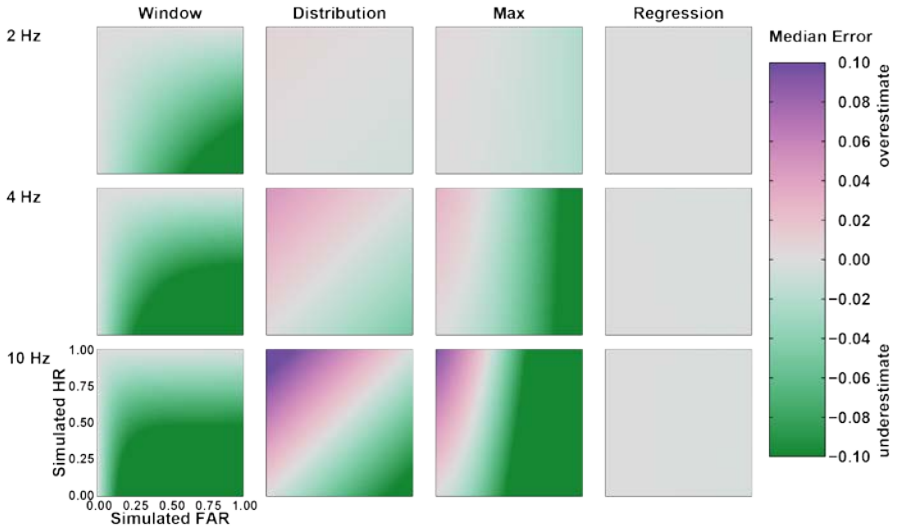

Fig. 5 FAR estimation error summary: panels show simulation results for estimation methods (columns) at a particular presentation rate (rows) when the true probability density function of the response times was known. Colors indicate difference between median estimate of FAR and simulated value of FAR clipped to an absolute value of 0.1. Within a panel, simulated FAR increases from left to right; simulated HR increases from bottom to top. All methods except regression have FAR estimation errors that clearly depend on the simulated HR and FAR, and overall magnitude of errors increases as presentation rate increases.

Table 1 HR estimate performance with accurate RT-PDF

	Presentation rate (Hz)					
	2		4		10	
Method	**err**	**std**	**err**	**std**	**err**	**std**
Window	0.280	0.016	0.447	0.012	0.518	0.008
Distribution	0.016	0.007	0.106	0.016	0.260	0.013
Max. attrib.	0.012	0.007	0.080	0.017	0.217	0.023
Regression	0.001	0.008	0.002	0.022	0.004	0.053

Notes: err = median absolute difference of each estimate from simulated values; std = standard deviation of estimates with median error subtracted.

Table 2 FAR estimate performance with accurate RT-PDF

	Presentation rate (Hz)					
	2		4		10	
Method	**err**	**std**	**err**	**std**	**err**	**std**
Window	0.053	0.003	0.085	0.002	0.099	0.002
Distribution	0.003	0.001	0.020	0.003	0.049	0.003
Max. attrib.	0.010	0.002	0.064	0.005	0.144	0.006
Regression	0.001	0.001	0.002	0.004	0.002	0.010

Notes: err = median absolute difference of each estimate from simulated values; std = standard deviation of estimates with median error subtracted.

3.1.7.3 Statistical Assessment of Bias in Estimation

To statistically assess the extent to which estimation errors depended on simulated HR, simulated FAR, and simulated presentation rate, HR estimation errors were first analyzed with a 4-way analysis of variance (ANOVA) with a categorical factor of estimation method (window, max, distribution, and regression) and continuous factors of presentation rate (2, 4, and 10 Hz), simulated HR, and simulated FAR (both ranging from 0 to 1 at 0.01 increments). Because the estimation method interacted with all other factors, individual ANOVAs were run for each method with factors' presentation rate, simulated HR, and simulated FAR. Results of method-specific analyses are in Table 3. In summary, all factors and interactions were statistically significant for the window method, with the 2 largest effects, measured with η^2, being the HR ($\eta^2 = 0.226$) and presentation rate ($\eta^2 = 0.225$). For the max method, all factors and interactions were statistically significant, with the interaction of presentation rate with HR ($\eta^2 = 0.240$) and the interaction of presentation rate with FAR ($\eta^2 = 0.355$) being the 2 largest effects. For the distribution method, all factors and interactions except the main effect of presentation rate and the interaction of HR with FAR were statistically significant, with the interaction of presentation rate with HR ($\eta^2 = 0.370$) and of presentation rate with FAR ($\eta^2 = 0.371$) having the largest effects. For the regression method,

HR, FAR, and the interaction of those with presentation rate as well as the 3-way interaction were statistically significant, but the effect sizes of all factors and interactions were less than 10^{-4}. This indicated that although the regression method's estimates do systematically depend on the presentation rate, HR, and FAR, the effects each account for less than 0.01 of a percent of the variance in the data. The statistical analysis on the FAR estimation errors produced similar results.

Table 3 Method-specific ANOVA

Source	d.f.	F	η^2	*p*-value
	Window method			
Rate	1	4.34×10^6	0.2245	0.0000
HR	1	4.37×10^6	0.2264	0.0000
FAR	1	2.08×10^5	0.0108	0.0000
Rate*HR	1	1.28×10^5	0.0066	0.0000
Rate*FAR	1	1.25×10^6	0.0645	0.0000
HR*FAR	1	1.36×10^6	0.0702	0.0000
Rate*HR*FAR	1	2.13×10^4	0.0011	0.0000
Error	7.65×10^6		0.3960	
	Max attribution method			
Rate	1	2.61×10^4	0.0009	0.0000
HR	1	1.53×10^6	0.0499	0.0000
FAR	1	2.56×10^6	0.0838	0.0000
Rate*HR	1	7.34×10^6	0.2400	0.0000
Rate*FAR	1	1.09×10^7	0.3554	0.0000
HR*FAR	1	5.83×10^4	0.0019	0.0000
Rate*HR*FAR	1	5.58×10^5	0.0182	0.0000
Error	7.65×10^6		0.2500	
	Distribution method			
Rate	1	1.26	0.0000	0.2620
HR	1	4.79×10^6	0.0718	0.0000
FAR	1	4.81×10^6	0.0720	0.0000
Rate*HR	1	2.47×10^7	0.3704	0.0000
Rate*FAR	1	2.48×10^7	0.3711	0.0000
HR*FAR	1	0.11	0.0000	0.7396
Rate*HR*FAR	1	6.71	0.0000	0.0096
Error	7.65×10^6		0.1146	
	Regression method			
Rate	1	0.00	0.0000	0.9565
HR	1	233.38	0.0000	0.0000
FAR	1	115.08	0.0000	0.0000
Rate*HR	1	101.94	0.0000	0.0000
Rate*FAR	1.00	190.046	0.0000	0.0000
HR*FAR	1.00	2.29	0.0000	0.1303
Rate*HR*FAR	1.00	15.40	0.0000	0.0001
Error	7.65×10^6		0.9999	

3.1.7.4 Simulations Run with Inaccurate RT-PDF Estimates

The second set of simulations used flat RT-PDF estimates to assess the performance of the RT-PDF-dependent methods when the estimated RT-PDF does not match the true RT-PDF. Numerical summaries for HR and FAR estimation are in Tables 4 and 5. Compared with results with the correct RT-PDF, the distribution and max methods both had larger errors over a larger range of HR and FAR when using the flat RT-PDF. The regression method's estimation errors increased somewhat by using the incorrect RT-PDF, but overall errors were smallest.

Table 4 HR estimate performance with flat RT-PDF

	Presentation rate (Hz)					
	2		4		10	
Method	**err**	**std**	**err**	**std**	**err**	**std**
Window	0.280	0.016	0.447	0.012	0.518	0.008
Distribution	0.190	0.014	0.299	0.009	0.333	0.005
Max attrib.	0.088	0.017	0.347	0.022	0.328	0.023
Regression	0.010	0.030	0.027	0.058	0.033	0.105

Notes: err = median absolute difference of each estimate from simulated values; std = standard deviation of estimates with median error subtracted.

Table 5 FAR estimate performance with flat RT-PDF

	Presentation rate (Hz)					
	2		4		10	
Method	**err**	**std**	**err**	**std**	**err**	**Std**
Window	0.280	0.016	0.447	0.012	0.518	0.008
Distribution	0.190	0.014	0.299	0.009	0.333	0.005
Max attrib.	0.088	0.017	0.347	0.022	0.328	0.023
Regression	0.010	0.030	0.027	0.058	0.033	0.105

Notes: err: median absolute difference of each estimate from simulated values; std = standard deviation of estimates with median error subtracted.

3.1.7.5 Analyzing Experimental Data

As an example of the effect of using different analytical methods on real data, the actual (rather than simulated) responses were analyzed. HR estimates are shown for each of the 15 subjects in Fig. 6. HRs were fairly high, ranging from 78.4% to 90.5% across subjects and estimation methods. A one-way repeated measures ANOVA revealed a significant effect of analysis method on HR estimate ($F(3,42) = 36.0$, $p = 1.1 \times 10^{-11}$, $\eta^2 = 0.131$). Follow-up paired comparisons (Bonferroni corrected) showed that the distribution ($M = 0.843$, $SE = 0.002$) and max ($M = 0.848$, $SE = 0.002$) estimates were not significantly different ($p = 0.25$), and the window ($M = 0.864$, $SE = 0.002$) and regression ($M = 0.864$, $SE = 0.002$) estimates were also not significantly different ($p = 1.0$), but both max and distribution estimates were significantly lower than both the window and regression estimates (all $p < 0.00001$).

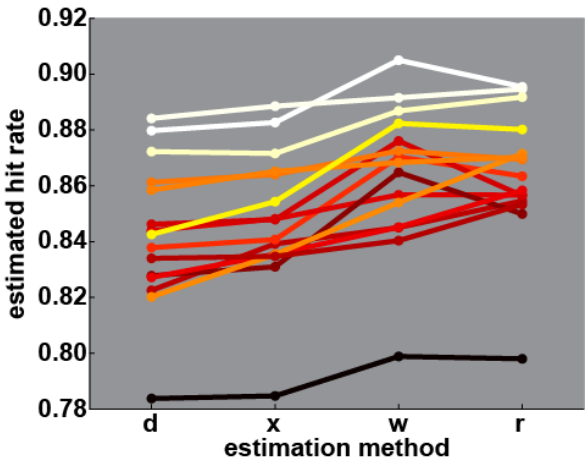

Fig. 6 Estimated HRs from experimental data with 4 different estimation methods: HR estimated from the response data from 15 subjects using the distribution (d), max (x), window (w), and regression (r) methods. Colors for individual subject data are based on estimates from the regression method to illustrate how relative ordering of subjects changes based on estimation method.

3.1.8 Summary and Discussion

The primary goal of these simulations was to test how well the proposed regression method for estimating HR and FAR in RSVP target detection tasks could recover the true simulated HR and FAR relative to established methods. The simulation results showed that the proposed regression method was more accurate than established methods, although accuracy comes at the cost of some precision.

The simulations comparing the performance of the 4 HR and FAR estimators revealed systematic errors in all 4 methods such that the error in HR and FAR estimates depended on some combination of the true value of the HR, FAR, and presentation rate, but the inaccuracy of the 4 methods were not equivalent.

The window method overestimates the HR as the true HR decreases and/or the true FAR increases. This can be understood as a result of the benefit-of-the-doubt approach this method represents. Any response within a window of a target is declared a hit by this method, so any false alarm that occurs in temporal proximity to a target might be incorrectly classified as a hit. Additionally, responses to targets that are slow enough to fall outside the 1-s window will be misclassified as misses. An important property of the window method demonstrated in the results here is that when the true FAR is very low, this method yields fairly accurate estimates of HR and FAR. This is because as the FAR approaches zero, the vast majority of responses will actually be hits, and the vast majority of hits should fall within the window and therefore be correctly classified by this method. This was illustrated in the "excellent" performance simulation (Fig. 3) in which the HR was slightly underestimated and the FAR was slightly overestimated.

Overall, the max and distribution methods made smaller errors in HR estimation than the window method (Table 1), although errors were relatively large in the range of HR and FAR that might be associated with good or excellent task performance (Fig. 3). These methods both had their lowest estimation errors when the simulated HR and FAR were similar. Because RSVP experiments typically report fairly high HR and low FAR, in practice both of these methods are expected to underestimate the HR and overestimate the FAR.

The regression method had lower estimation error than the other 3 methods, and the errors do not depend strongly on the true values of HR and FAR. The distribution method makes systematic errors that depend strongly on the true HR, FAR, and presentation rate (Table 3), and the regression method attempts to correct for those errors by accounting for how errors contribute to the expected value of the apportionment to any given stimulus using linear regression. The statistical analysis of the estimation errors of the regression method revealed a reliable effect of the interaction of FAR with presentation rate, but the effect size was less than 10^{-4}. The absence of nontrivial linear effects revealed in the ANOVA is evidence that the linear regression method accomplished its goal. Nonlinear effects could potentially affect the estimation error of the regression method, but given the small overall estimation error of the regression method (Tables 1 and 2), any such effects do not appear to have a major impact, at least under the conditions simulated here.

The presentation rate had a sizeable impact on estimate accuracy in all of the estimation methods except the regression method, although the precision of the regression method's estimates decreased as the presentation rate increased. The increases in estimation error can be understood as a consequence of the increasing ambiguity of which stimulus elicited a particular response. Although such a slow rate was not tested here, clearly if the stimuli are spaced far enough apart, then errors in response assignment would be very rare. As more stimuli fall into a temporal range of plausibly causing a response, the harder it will be to correctly assign that response to a stimulus.

One potential caveat to the apparent success of the regression method is that in our simulations, the regression method was provided with the true probability density function for response times (RT-PDF). In practical use, the RT-PDF would have to be estimated from the available data. For completeness, simulations included true HRs that were low or zero. In those situations, estimating an RT-PDF would be difficult or impossible, so in our second set of simulations we provided all of the analytical methods with a highly incorrect, uniform RT-PDF (Tables 4 and 5). Having such a poor estimate of the RT-PDF did not obliterate the RT-PDF-dependent methods, although the performance of those methods did drop

somewhat. Based on this result, it seems that even if estimation of the RT-PDF is poor, the regression method may still be recommended.

An assumption of the regression method is that responses to different stimuli are independent. Strictly, this assumption is incorrect for 2 reasons. First, the method assumes that it is possible for 2 responses to occur at the same time (e.g., a slow response to an earlier stimulus occurs simultaneously with a fast response to a later stimulus), but in practice there are limits to how quickly a person can press a button twice. This first assumption was in fact violated in the simulations run here because in the rare event that multiple responses occurred at the same time, those responses were conflated into a single response. The chance of response collisions increases as the number of overall responses increases, and this would be most prevalent at fast presentation rates with high FARs. It might also explain the small but significant interaction of FAR with presentation rate that impacted the estimation error of the regression method.

Second, humans typically fail to perceive images that fall within a short window of time starting shortly after a target image. This phenomenon is called the attentional blink.[53,54] This could temporarily lower the HR and/or the FAR by reducing the probability of responding for a short time after each response. There was no modeling of the attentional blink in the simulations done here, so its impact on any of the estimation methods here cannot be assessed.

To illustrate the impact the choice of HR/FAR estimation method might have in an experimental setting, behavioral results from a target-detection experiment were analyzed using the 4 methods tested in simulations. The impact of analysis method on the overall HR and FAR estimates was statistically significant, and the effect of analysis method was consistent with the simulation results of good performance overall. Qualitatively, this provides support for the validity of our simulations. However, for some individuals, the regression method estimated a somewhat higher HR compared with the window method (Fig. 6). Inspection of the responses from the subjects for whom the regression method had a higher estimate than the window method revealed that these subjects appeared to occasionally respond twice within a 500-ms span (corresponding to the inter-stimulus interval). If a single target image elicits 2 responses, the window method calls one a hit and the other a false alarm, so double-responding would not inflate the HR estimate. The regression method, however, does not have special handling of double responses and they could inflate the HR estimate. Based on these data, we cannot know if these responses are examples of nonindependence. It could be that the subjects inadvertently pressed the response button twice after seeing a target image, or it could be that the 2 BPes were intended as responses to consecutive images.

Based on its better estimation of HR and FAR, the regression method proposed here would seem the best choice when estimating the HR and FAR is the primary interest. If the FAR is known to be essentially equal to 0, the window method may have an advantage because the window method is somewhat simpler to implement and is more precise with faster presentation rates. In the more general case in which the FAR may be nonnegligible and a fast presentation rate is used, the regression method is likely to provide the most accurate estimates of HR and FAR.

Much like machine-learning classifiers, optimally incorporating human input into a heterogeneous target detection system requires a training set of labeled inputs to learn a confidence metric for the human detector. The method developed under this Director's Strategic Initiative (DSI) affords better estimates of human performance than existing methods, which will allow optimal integration of human input into a human–autonomy team for target image detection.

3.1.9 Future Work/Transitions

Although the focus of this method is on target-detection accuracy in the RSVP paradigm, many related projects focus on using some physiological measure to enable a brain–computer interface (BCI) for target detection.[43,44,55,56] Electroencephalography (EEG)-based classification can sometimes classify images correctly even when the behavioral response was incorrect.[57,58]

One transition of this work is toward integrating human input into heterogeneous teams of human and automated image detectors. To that end, software implementing the method described here was provided to a team operating under a complementary research effort that is engineering an integrated heterogeneous target-detection system.

Recording EEG from a participant engaged in a RSVP target-detection task affords an opportunity to, in parallel with a direct BCI, identify physiological features that relate to changes in task performance. A future transition of this work will be to develop predictive models that accomplish this. Initial efforts have shown that the mean amplitude of alpha-band oscillations (8–12 Hz periodic activity in the ongoing EEG) within some window of time are correlated with the participants' overall HR for that same window of time. This relationship shows between-subjects variability, however, such that the correlation between alpha and performance takes on a range of both positive and negative values. However, the magnitude of the correlation between short-time alpha amplitude and short-time performance changes is itself negatively correlated with the overall performance of the participants (Fig. 7). So it appears that changes in performance may be predicted from alpha-band activity in some individuals, but those individuals have overall

worse performance. Future work will determine if this alpha/performance relationship persists across multiple days, or if it varies situationally. Should this relationship be found to have some stability, it could be used to identify consistent high performers as well as indicate when more variable performers are likely to be at their best.

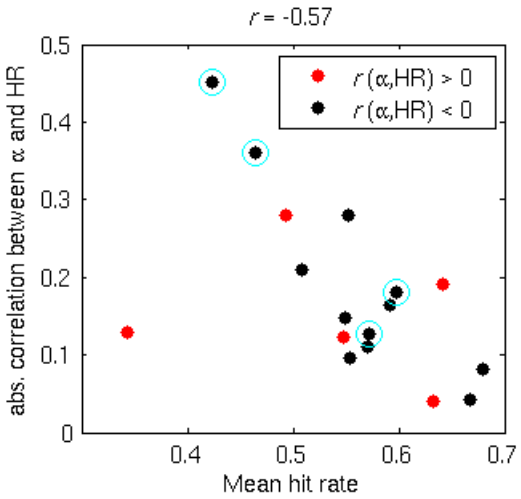

Fig. 7 Correlation between short-time alpha-band activity and short-time HR is predictive of overall performance. Participants (N = 17) viewed images of a cluttered office and pressed a button when they saw target objects (e.g., a chair). HR and alpha-band EEG were estimated from a sliding 3-min window. The absolute value of the correlation between these 2 estimates is plotted on the ordinate; HR estimates for each subject from the entire time course of the experiment are on the abscissa.

3.2 Developing Confidence in Human Performance for RSVP Performance on Complex Tasks

Previous studies using RSVP tasks for rapid target detection have primarily focused on the 2-class discrimination problem of detecting target images within a set of distractor images.[40,47,59–67] However, in many real-world environments there is likely to be a subset of distractor stimuli that share physical and semantic features with the target stimuli (e.g., consider a nontarget elk versus a target deer in a dense ensemble of forest imagery). While event-related potentials (ERPs) studies have analyzed the neural features evoked by rare nontargets within a series of rare targets and frequent background distractors using simple classes of stimuli (e.g., letters and colored shapes),[68] it is unknown if similar effects occur in complex imagery more similar to real-world settings. Moreover, little research has been done to evaluate how current neural-based classification algorithms perform when 2 infrequent classes of stimuli with the same features (i.e., target and nontarget) are presented in a sequence of frequently occurring distractor images. It is possible that many classification algorithms used for RSVP target-detection studies are sensitive to

neural features primarily associated with the detection of infrequent stimuli rather than target detection/recognition, resulting in drastically reduced performance.

The RSVP-based image triage process uses a measure of confidence in the classifier through the probability score as a means of quantifying the certainty of a decision. That is, the probability that a particular image is a target provides information regarding the likelihood a target was presented. The importance of confidence in systems with low signal-to-noise properties has long been understood in decision theory[30–32] and control communities[4,33] and peripherally exists in current instantiations of image triage BCIs.[59,60,69,70] Additional uses of confidence measures in BCIs are demonstrated through the rejection of particular trials from analysis or the use of algorithms for the removal of artifacts. Thus, while the use of confidence measures for target-detection BCIs is not new, previous studies have not explicitly described their methods for deriving the confidence metric and have not quantified the accuracy of their confidence estimates or the unique contribution of confidence itself.

This study explores how current RSVP-based BCI technologies may function in more-complex task environments by adding infrequent nontarget images that are not task-relevant but are physically and semantically similar to targets to presentations with rare targets and frequent background distractors. In the first half of this report, we examine participants' ability to detect targets under 2 conditions: when targets are the only infrequent image class presented and when the targets are presented with infrequent nontargets in a standard RSVP task. Our analysis encompasses behavior, averaged ERPs, and single-trial classification of EEG data. The results demonstrate that both behavioral and single-trial classification performance of target images decline with the introduction of rare visually similar nontarget stimuli. We also examine the effects of using trial-by-trial confidence measures derived from the relationship between individual classifier outputs and the discriminating threshold between targets and nontargets to mitigate the drop in classifier performance. These results provide a unique perspective into how methods for EEG classification of visual imagery may perform in more-complex scenarios and the importance of incorporating confidence.

3.2.1 Methods

3.2.1.1 Participants

Eighteen participants volunteered for the current study. Participants reported normal or corrected-to-normal vision and no history of neurological problems. Due to excessive artifacts in the EEG data, one participant was excluded from analysis.

The resulting 17 participants had an average age of 34.9 years, 14 were male, and all participants were right-handed with the exception of one left-handed male.

The voluntary, fully informed consent of the persons used in this research was obtained as required by federal and Army regulations.[49,50] The investigator adhered to Army policies for the protection of human subjects.[50]

3.2.1.2 Stimuli and Procedure

Participants were seated 75 cm from a monitor and viewed a series of images from a simulated desert–metropolitan environment in an RSVP paradigm (Fig. 1). Images (960 × 600 pixels, 96 dpi, subtending 36.3° × 22.5°) were presented using E-prime software for 500 ms (2 Hz) with no inter-stimulus interval.

Data were analyzed from 2 conditions for all participants: Target Only (TO) and Target and Non Target (TN). The TO condition contained only background distractors (background scenes of a desert–metropolitan environment) and target images (background scenes with a person carrying a weapon). The TN condition contained nontarget distractor stimuli (background scene with a person without a weapon) along with both background and target stimuli (Fig. 8). Target stimuli (both TN and TO conditions) and nontarget distractor stimuli (TN condition only) were never presented back to back. At least 2 background stimuli were required to follow any target or nontarget stimulus to avoid issues with the attentional blink.[71,72] In both the TO and TN conditions, participants were instructed to press a button on a serial response box as rapidly and accurately as possible with their dominant index finger when they detected a target. Participants were also instructed to silently count the number of targets they detected and report this number at the end of each block.

Fig. 8 RSVP task and stimuli in the current experiment: participants required to detect target images while ignoring nontargets and background distractors.

Each condition contained 6 blocks of RSVP image sequences. Each block was a 2-min image sequence in the TO condition and a 2-min-and-14-s sequence in the

TN condition. The interblock rest period was self-paced after a mandatory 10-s pause to report the target count. Each block began with a visual 5-s countdown presented at the center of the display. Participants were told to fixate toward the center of the display, as all target and nontarget stimuli appeared within 6.5° of the image center and would not appear on top of or occluded by buildings and trees or in windows. Block order was counterbalanced across participants. The individual blocks served to break up the RSVP presentation and allow subjects to periodically rest. Thus, data from the 6 blocks within each condition were concatenated and analyzed as a whole.

The target-to-distractor ratio was 1:20 in the TO condition and 1:14 in the TN condition. The nontarget-to-distractor ratio in the TN condition was also 1:14. Participants were not aware of stimuli contingencies. Participants were given one block of practice on each RSVP stimulus condition and were required to correctly report at least 75% of targets to begin the experiment. All participants needed only one practice block in each condition to satisfy this requirement.

3.2.1.3 EEG Recording and Preprocessing

Electrophysiological recordings were digitally sampled at 1024 Hz from 64 scalp electrodes arranged in a 10–10 montage using a BioSemi Active Two system (Amsterdam, Netherlands). Impedances were kept below 25 kΩ. External leads were placed on the outer canthus of each eye and above and below the right orbital fossa to record electro-oculogram (EOG). Continuous EEG data were preprocessed using EEGLAB software.[73] The EEG data were referenced to the average of the left and right earlobes, decimated to 512 Hz, and digitally filtered 0.1–50.0 Hz.

Gross artifacts were removed through visual inspection of the continuous EEG data. Sections marked as artifacts were excised from the data. Subsequently, independent component analysis (ICA)[74] was run. Independent components related to eye movements or muscle activity were manually identified and removed. The time series data resulting from the ICA-based cleaning was used for all further analyses.

For single-trial classification, the signal was first bandpass filtered (Butterworth filter of order 4) with cutoff frequencies at 1 and 10.66 Hz and then downsampled to 32 Hz. This new sampling rate was chosen based on the sampling frequency used by the winning team of the competition in the 2010 IEEE Workshop on Machine Learning for Signal Processing.[75]

3.2.1.4 Behavioral Analysis

To quantify the behavioral performance, any BP that occurred between 200 and 1000 ms after a target or nontarget stimulus was attributed to that trial. Button

presses attributed to target trials were counted as hits and all others as false positives (FPs). Reaction times were calculated as the time between stimulus presentation and BP.

Hits (*Hit*), misses (*Miss*), correct rejects (*CorrectReject*), and *FP*s were calculated for each subject. The correct rejects and false alarms were calculated separately for nontargets and distractor trials in order to investigate the effect of adding the nontarget stimuli to the behavioral performance. These values were used to calculate *d´* (d-prime), an index of accuracy that accounts for response bias,[76] for each subject, as follows:

$$HR = \frac{Hit}{Hit+Miss} \qquad FPR = \frac{FP}{FP+CorrectReject}. \tag{7}$$

$$d' = Z(HR) - Z(FPR), \tag{8}$$

where the function $Z(p)$, $p \in [0,1]$, is the inverse of the cumulative Gaussian distribution.

3.2.1.5 ERP Analysis

ERP data were processed and analyzed using ERPLAB.[77] Artifact-free data were epoched [–500, 1000] ms around stimulus onset and binned according to the experimental condition. ERPs were baseline corrected by subtracting the mean of the activity of each channel from [–500, 0] ms from the epoched data. Only hits and correct rejections were included in the ERP analysis. ERPs were calculated for each stimulus type (background distractors, targets, and nontargets). P3 amplitude (400–800 ms) was separately calculated for each subject in each experimental condition at electrode Pz. The time segment analyzed was chosen based on the grand-average target ERP waveforms, which showed the maximum P3 amplitude occurring over electrode Pz 400–800 ms poststimulus.

3.2.1.6 Single-Trial Classification

To quantify the effects of adding rare, target-like nontarget stimuli at the single-trial level, EEG data were epoched to [0, 1000] ms, timelocked to stimulus onset, spatial filtered using xDAWN[78], and classified with Bayesian linear discriminant analysis[79] (collectively referred to as XD+BLDA).[61,78,80,81]

XD+BLDA

The xDAWN algorithm is a spatial filtering algorithm that identifies a linear combination of the raw neural signals that maximizes the signal-to-noise ratio

between targets and nontargets. Let $U \in \mathbb{R}^{N_s x N_f}$ be the spatial filters, where N_s is the total number of sensors and N_f is the number of spatial filters. The signal after spatial filtering is defined by $X_{filt} = XU$ where $X \in \mathbb{R}^{N_t x N_s}$ is the recorded signal and N_t is the number of sampling points. The expected waveform is considered spatially stable over time for the spatial dimension reduction step.

In this framework, an algebraic model of the enhanced signals XU is composed of 3 terms: the ERPs evoked by the targets (D_1A_1), a response common to all stimuli (D_2A_2), and the residual noise (H), which are spatially filtered with U.

$$XU = (D_1A_1 + D_2A_2 + H)U. \quad (9)$$

D_1 and D_2 are 2 real Toeplitz matrices of size N_t x N_1 and N_t x N_2, respectively. D_1 has its first column elements set to zero except for those that correspond to a target onset, which are set to 1. For D_2, its first column elements are set to zero except for those that correspond to all stimulus onsets. A_1 and A_2 are 2 real matrices of size $N_1 \times N_s$ and $N_2 \times N_s$, respectively. A_1 represents the prototypical ERP in response to targets, and A_2 represents the prototypical ERP in response to all stimuli. N_1 and N_2 are the number of sampling points representing the target and superimposed evoked potentials, respectively. H is a real matrix of size $N_t \times N_s$.

Let us define spatial filters U that maximize the signal to signal plus noise ratio (SSNR) as follows:

$$SSNR(U) = \frac{Tr(U^T \hat{A}_1^T D_1^T D_1 \hat{A}_1 U)}{Tr(U^T X^T X U)}, \quad (10)$$

where $\hat{A}_1$ corresponds to the least mean square estimation of A_1.

$$\hat{A} = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \end{bmatrix} = ([D_1; D_2]^T [D_1; D_2])^{-1} [D_1; D_2]^T X, \quad (11)$$

where $[D_1;D_2]$ is a matrix of size $N_t * (N_1+N_2)$ obtained by concatenation of D_1 and D_2. Spatial filters are obtained through the Rayleigh quotient by maximizing the SSNR.[78] The result of this process provides N_f spatial filters, which are ranked in terms of their SSNR.

Eight spatial filters ($N_f = 8$) are then used as input to a BLDA classifier. The input vector is obtained by concatenating the N_f time-course signals across the resulting spatial filters. The BLDA classifier was selected because it is relatively robust to noise in the training data.[79,82]

Confidence

Confidence measures were derived to identify the reliability of the classification made for each trial. A simple measure, the distance of the classifier score to the discriminating boundary, was used as confidence, as follows:

$$Conf = \begin{cases} \frac{Score-\kappa}{\max(Score)-\kappa} & Score > \kappa \\ \frac{Score-\kappa}{\min(Score)-\kappa} & Score \leq \kappa \end{cases}, \quad (12)$$

where *Score* is the score produced by the XD+BLDA classification on a single trial. The classifier score represents a projection from the feature space down to the decision space that maximally separates the 2 classes. κ is the threshold established through XD+BLDA for discriminating targets from nontarget and background distractor stimuli. Max(*Score*) and min(*Score*) are the maximum and minimum scores over the entire training set.

Performance Evaluation

The effect of including the visually similar nontarget stimuli in the RSVP paradigm on classifier performance was explored by comparing the classifier performance across the TO and TN conditions 3 distinct discriminations. First, target stimuli were discriminated from background distractor stimuli in the TO condition. This discrimination represents the baseline RSVP paradigm with only 2 types of stimuli. Next, we discriminated target stimuli from background distractor stimuli in the TN condition, omitting the nontarget stimuli. Then we discriminated target stimuli from both nontarget and background distractor stimuli in the TN condition.

For each discrimination, classifier performance was evaluated using a nested 10-fold cross validation with 80% of the data used to train the spatial filter and classifier, 10% of the data used to test the classifier and establish discrimination thresholds, and the remaining 10% of the data were used as an independent validation set on which to apply the trained classifier and thresholds. This process was repeated 10 times, such that each contiguous 10% slice of data was used as the final validation set. Performance was evaluated based on the area under the receiver operating characteristic (ROC) curve (Az)[83] and misclassification rate in the final validation sets.

Misclassification rates were derived based on a discrimination threshold that maximizes the difference between the true positive rate and the FP rate from the classifier scores in the training set and then applying this threshold to the classifier scores in the validation set. Both Az and misclassification rates were also used to quantify the accuracy of the confidence measures presented here. To do so, a

threshold for dividing the data into high-confidence and low-confidence subsets was varied from 0% to 90% in steps of 10%. A confidence threshold of 0% meant that 0% of the data was included in the low-confidence subset, and all of the data was included in the high-confidence subset. A confidence threshold of 90% indicated that 90% of the data was included in the low-confidence subset and 10% of the data was included in the high-confidence subset. For each confidence threshold in this range, the Az and misclassification rates of the high-confidence subset were measured. Using these metrics, confidence values that accurately represent the reliability of performance should increase Az and decrease misclassification rates as the confidence threshold is raised.

Mitigation Strategies

The utility of applying confidence measures was further demonstrated by quantifying the improvement in image-labeling accuracy when the estimated confidence was used to trigger a corrective action. This study simulated a simple mitigation strategy where trials above the confidence threshold were classified using the neural classifier (NC) and trials below the confidence threshold were manually labeled by the participant. For the purpose of this simulation, we assume that a human participant given unlimited time to label the image will attain 100% accuracy; thus, the manually labeled trials were set to the actual image labels. The classification performance using this simulated mitigation strategy was evaluated using Az and misclassification rates for each stimulus class.

3.2.2 Results

Results across the behavioral, ERP, and single-trial classification analyses demonstrated that adding sparse, visually similar, nontarget images made it more difficult for participants to identity target images.

3.2.2.1 Behavior

Behavioral performance was characterized by comparing the error rate by stimulus type, reaction time, and d´ across the TO and TN conditions (Fig. 9). Across all 3 measures, behavioral performance declined when nontargets were included. Adding nontargets more than doubled the average error rate for target stimuli (difference significant, Wilcoxon signed rank test, $p < 0.01$, Fig. 9A). Reaction times obtained from correct target trials were significantly faster in the TO condition (median RT of 514.67 ms) than in the TN condition (median RT of 602.82 ms) (Wilcoxon signed rank test, $p < 0.001$, Fig. 9B). D-prime analysis showed that target discrimination performance was significantly better for TO trials

(median d´ of 4.25) over TN trials (median d´ of 3.49) (Wilcoxon signed rank test, $p < 0.01$, Fig. 9C).

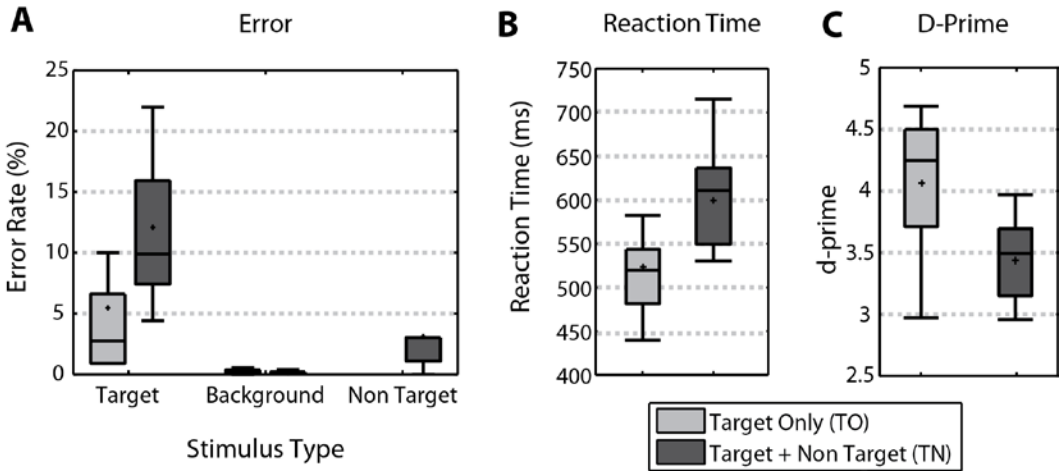

Fig. 9 Behavioral performance: Panel A shows error rates for each stimulus type for both TO (light gray) and TN (dark gray) conditions; Panel B shows target reaction time for both conditions; Panel C shows d´ measures for both. Error bars show highest and lowest data point within 1.5 times the interquartile range of upper and lower quartiles, respectively. Within each box, crosses indicate mean values and horizontal lines indicate median values.

3.2.2.2 ERP Analysis

Statistical comparisons of grand-average ERP waveforms demonstrated that ERPs were significantly different across stimulus type, with visually similar nontargets generating ERPs with amplitudes between those of target stimuli and background distracters. In addition, ERPs for background distractor and target stimuli were not significantly different across the TO and TN conditions. A one-way ANOVA was used to analyze the mean amplitude (400–800 ms) from electrode Pz with stimulus (background distractor, target, and nontarget) as a main factor. There was a main effect for stimulus in the TO condition, ($F(1,16) = 111.34$, $p < 0.001$), indicating a significantly larger P3 amplitude for targets (mean amplitude: 13.66 µV) relative to background distractors (mean amplitude: –0.44 µV , Fig. 3A). A main effect was also obtained in the TN condition ($F(2,32) = 83.01$, $p < 0.001$). Subsequent multiple comparison tests using the Tukey–Kramer method showed that amplitudes from background distractors, targets, and nontargets were all significantly different from each other (Fig. 3B). A 2-way ANOVA was run with the factors of condition (TO or TN) and stimulus (distractor or target) to assess any differences between target P3 amplitude in the 2 conditions. There was a main effect of stimulus ($F(1,16) = 344.33$, $p < 0.001$) but no main effect for condition ($F(1,16) = 0.001$, $p = 0.978$) or interaction ($F(1,16) = 0.002$, $p = 0.964$), indicating that both the background distractor and target activity was similar between the TO and TN conditions and

that there were significant differences between background distractor and target activity in both the TO and TN conditions (Fig. 10).

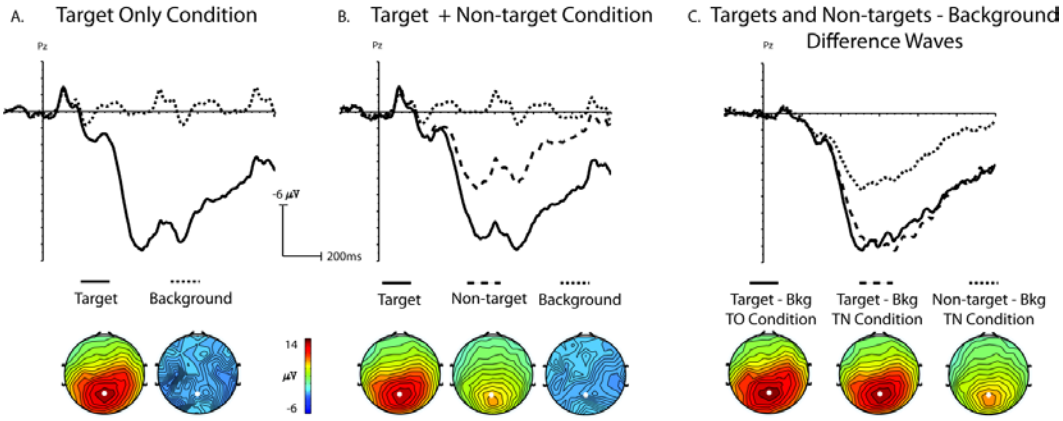

Fig. 10 Grand-average ERP waveforms at electrode Pz and topographic voltage maps (400–800 ms); white dot indicates location of electrode Pz. Panel A shows grand-average ERP waveforms and topographic maps to target and background distractor stimuli in the TO condition; Panel B shows grand-average ERP waveforms and topographic maps to target, nontarget, and background distractor stimuli in the TN condition; Panel C shows difference waves created by subtracting background distractor from targets in TO condition and the background distractor from targets and nontargets in TN.

3.2.2.3 Single-Trial Detection

Overall classification performance declines when visually similar nontarget stimuli are present in the RSVP stream (Fig. 11). The TO condition represents the baseline RSVP discrimination of target versus background distractor. The classifier was highly accurate in this condition, producing average Az values greater than 0.97. When targets are discriminated from background distractor stimuli in the TN condition (ignoring nontarget stimuli) performance is not significantly different (Wilcoxon signed rank test, $p = 0.06$). However, when nontarget stimuli are included in the discrimination, performance is significantly worse than when they were not included (Wilcoxon signed rank test, $p < 0.001$).

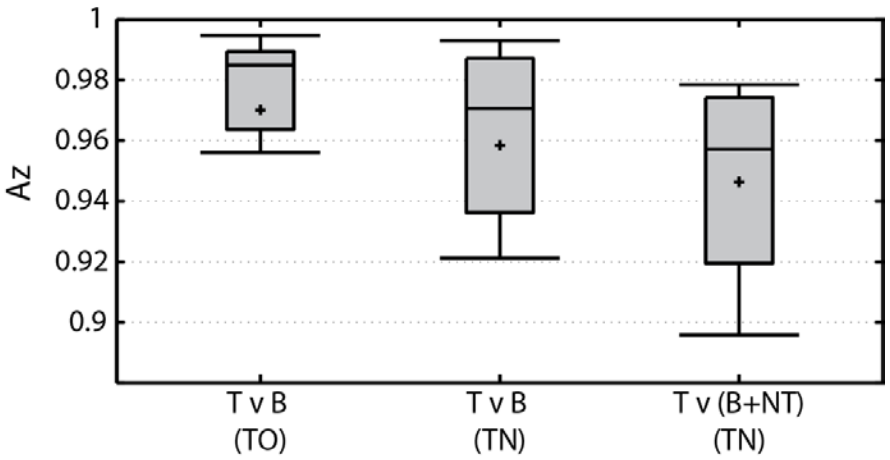

Fig. 11 Overall classification performance under various conditions: left, target vs. background distractor (T v B) discrimination performance in TO condition; middle, target vs. background distractor (T v B) discrimination performance in TN; right, target vs. both background distractor and nontarget [T v (B+NT)] discrimination performance in TN.

In addition to the Az measure, the classifier performance was also measured by quantifying the misclassification rate for each stimulus type (Fig. 12). Again, we focused on the same 3 discriminations: target versus background distractor in the TO condition (Fig. 12A), target versus background distractor in the TN condition (Fig. 12B), and target versus both nontarget and background distractor stimuli in the TN condition (Fig. 12C). In the baseline TO condition, misclassification rates were below 10% for both target and background distractor stimuli. This level of accuracy would be expected given the high Az levels achieved by in this condition (see Fig. 11).

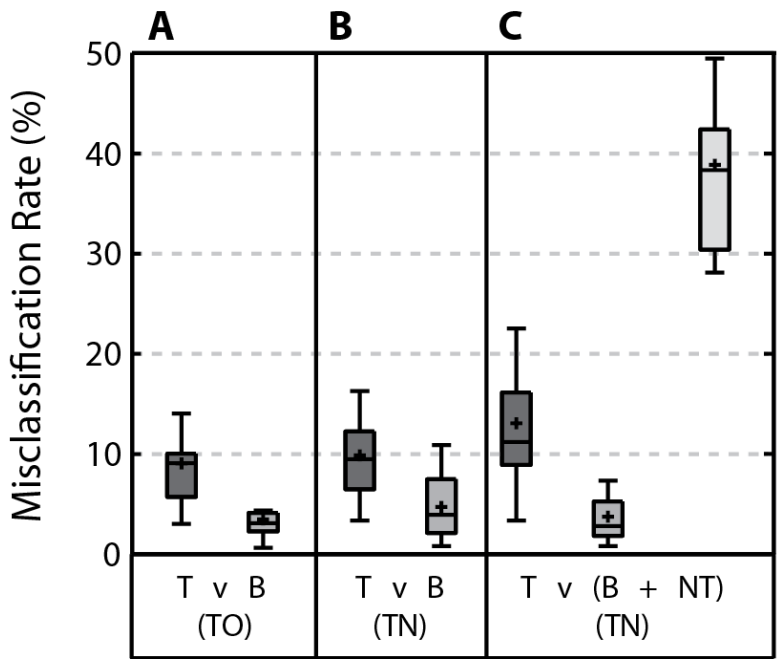

Fig. 12 Misclassification rate for each stimulus type for each discrimination when threshold was calculated based on classifier scores from training set. Panel A shows the misclassification rate for target (T) and background distractor (B) stimuli in the TO condition; Panel B shows the misclassification rate for T and B stimuli in the TN when targets are discriminated from B only; Panel C shows the misclassification rate for T, B, and nontarget (NT) stimuli in the TN when targets are discriminated from both B and NT stimuli. Error bars show highest and lowest data.

Moving from the TO condition to the TN condition resulted in no significant change in misclassification rates when discriminating target stimuli from background distractor stimuli (Wilcoxon signed rank test, $p = 0.23$ and $p = 0.07$ for target and background distractor stimuli respectively). Including nontarget stimuli in the discrimination increased misclassification rates for target stimuli (Wilcoxon signed rank test, $p = 0.01$) and resulted in an exceptionally high misclassification rate for nontarget stimuli (38.84% ± 8.71%). Misclassification rates for background distractor stimuli were slightly, yet significantly, reduced with the addition of the nontarget stimulus (Wilcoxon signed rank test, $p = 0.049$).

The increase in misclassification rates in the nontarget condition is potentially problematic for many real-world applications of this technology where users will encounter instances of nontarget stimuli that share the same physical and semantic features as target stimuli. To address this issue, we explored applying confidence measures to the classifier outputs as a means to mitigate the misclassification rate (Figs. 13 and 14).

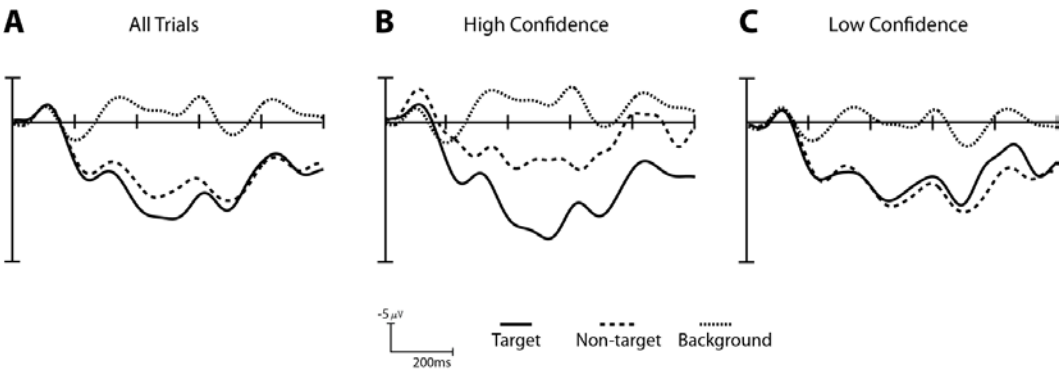

Fig. 13 Confidence ERPs for Subject S10: Panel A, ERPs across all trials; Panel B, ERPs for the high-confidence trials (e.g., top 25% trials when sorted by confidence); Panel C, ERPs for low-confidence trial (e.g., bottom 25% trials when sorted by confidence). The difference between high- and low-confidence waveform for all 3 stimulus categories is statistically significant (Wilcoxon signed rank test corrected for multiple comparisons using False Discovery Rate $p < 0.001$). High-confidence trials show greater separation between target and nontarget trials compared with low-confidence trials.

Nontarget ERPs from high-confidence trials are more readily distinguished from target ERPs than in low-confidence trials, as shown for subject S10 in Fig. 13. Here, high and low-confidence trials are defined as the top 25% and bottom 25%, respectively. Trials labeled with high confidence showed greater separation between target trials and both nontarget and background distractor trials than trials with low confidence. A Wilcoxon signed rank test (corrected for multiple comparisons using False Discovery Rate)[84,85] shows that the difference between the high- and low-confidence waveform for all 3 stimulus categories is statistically significant ($p < 0.001$). When this analysis is extended across all participants, 14 out of 16 participants show significant differences between high- and low-confidence trials for all 3 stimulus categories (p-values corrected for multiple comparisons using False Discovery Rate, $q = 0.05$). All participants had significant differences between high- and low-confidence stimuli for at least 2 of the 3 stimulus categories. A similar analysis was carried out to compare behavioral performance between high- and low-confidence trials (as defined by the classifier), but no significant difference was found.

Overall, nontarget stimuli have lower confidence than the target or background distractor stimuli (0.442 ± 0.0057, 0.5751 ± 0.0014, 0.3051 ± 0.0057 mean ± standard error for target, background, and nontarget stimuli, respectively; Fig. 14A). For each participant, a one-way repeated measures ANOVA was used to analyze the confidence attributed to each stimulus type. When p-values are corrected for multiple comparisons using False Discovery Rate analysis,[84,85] all 16 participants showed a significant effect for stimulus type ($q < 0.05$). Across all participants, the multiple comparisons analysis showed that the confidence

attributed to nontarget trials was significantly lower than the confidence attributed to both background distractor and target trials for all participants. Additionally, confidence values for target stimuli were less than those for background distractor stimuli.

The use of confidence measures also had a significant effect on classification performance. Figure 14B shows the area under the Az for classification performance for all trials as a function of confidence thresholds. As the confidence threshold is raised from the minimum to a value that matches the 90th percentile of confidence values for each subject, the average Az value across all participants increases to a nearly perfect classification (solid line in Fig. 14B). This improvement is further evidenced through the change in misclassification rates for each of the stimulus classes, as shown in Fig. 14C (solid lines). As the confidence threshold increases, misclassification rates for the target and background distractor stimuli fall to nearly zero. However, nontarget stimuli maintain a high level of misclassification regardless of confidence level. The improved performance obtained by raising the confidence threshold comes at the cost of ignoring portions of the dataset. The amount of data remaining for each stimulus class for increasing confidence thresholds is shown in Fig. 14D.

Alternatively, however, instead of simply ignoring trials that fall below a confidence threshold, one might instead choose to seek alternative methods for classification. A simple example of an alternative method would be to manually label those images where the NCs failed to produce a highly confident outcome. The performance of such a system improves the overall classification accuracy, as shown in the dashed line in Fig. 14B, at the expense of the extra time needed to manually label images. The performance improvement through the manual labeling process is further evidenced through the reduction of misclassification rates for each stimulus class (Fig. 14C, dashed lines). For background and nontarget stimuli, the difference between the neural classification alone and the neural classification combined with manual labeling is significant for all confidence thresholds above 0% (Wilcoxon signed rank test, $p < 0.001$ for both classes; p-values were also corrected for multiple comparisons through False Discovery Rate with $q < 0.05$). For target stimuli, the difference is significant for all confidence thresholds above 0% and less than 90% (Wilcoxon signed rank test $p < 0.001$ for both classes, p-values were also corrected for multiple comparisons through False Discovery Rate with $q < 0.05$).

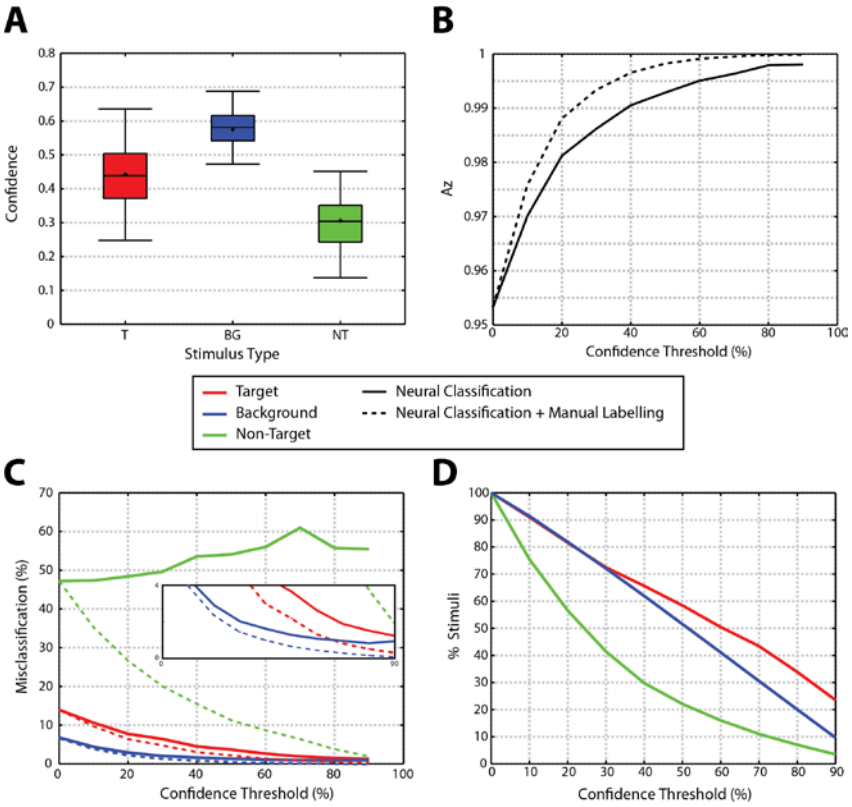

Fig. 14 Confidence: Panel A, confidence levels by stimulus type. Panel B, Az for trials as a function of confidence threshold (solid line shows the Az for trials exceeding confidence threshold given; dashed line shows Az when trials below confidence threshold are manually labeled while trials above threshold are labeled through neural classification; in both cases, as confidence increases, Az increases). Panel C, misclassification rates for trials exceeding a given confidence threshold. (Solid lines show misclassification rates for neural classification only. As confidence increases, misclassification rates for target and background distractor stimuli fall to nearly zero. Nontarget misclassification rates remain high regardless of confidence levels. Dashed lines show misclassification rates when trials below threshold are manually labeled, while trials above threshold use neural classification. Misclassification rates for all 3 stimulus classes are reduced through manual labeling process. Inset zooms in on lower portion of the graph, highlighting decrease in misclassification rates for target and background stimuli.) Panel D, percentage of trials exceeding a given confidence threshold.

3.2.3 Discussion

Prior work by many groups[40,47,59–67] has demonstrated the effectiveness of using single-trial classification to detect targets in RSVP; however, little of this work explicitly examined how feature similarity between target and nontarget stimuli effected target-detection accuracy. We addressed this concern in the present study by introducing a more-realistic situation where target and nontarget stimuli, though each occurred infrequently, shared both physical and semantic features but only targets were task-relevant. We evaluated the impact of this manipulation on behavior, ERPs, and single-trial classification of the evoked neural response.

Results across the behavioral, ERP, and single-trial classification analyses demonstrated that adding sparse, visually similar, nontarget images made it more difficult for participants to identity target images and more difficult to classify images from neural data.

3.2.3.1 Confidence

Previous studies using RSVP-based neural technologies for image triage applications[59,60,69,70] have employed statistical methods to identify a subset of trials most likely to be target images. As an extension of this previous work, we employed a confidence-based approach in an offline simulation to mitigate the drop in performance that occurred when nontargets were included in the RSVP stream.

Confidence measures derived from the classifier score were used to sort the dataset based on likelihood of correct classification. A comparison of the ERPs and single-trial classification performance showed significant differences between the high- and low-confidence trials. The ERP analysis showed that high-confidence target trials were more separate from the nontarget and background distractor trials than low-confidence T trials. This increased separation led to an improved classification performance for high-confidence trials. Specifically, Fig. 14B shows that as we remove the lower-confidence trials from the performance analysis, classification accuracy improves.

However, the use of a distance from threshold method for establishing confidence, as was done here, has been shown to be less than ideal in previous studies.[86] Employing more-robust confidence measures (for example, a density-based estimation method in the learned feature space) will likely further improve performance. Additionally, our confidence measures used only information from the classifier scores; however, there is potentially a large amount of information in a variety of sources that could further improve the estimate of confidence in a given decision (e.g., data from multiple sensor modalities, individual skill level/expertise, and sleep history). We envision that an accurate estimate of confidence in a particular decision (e.g., target versus nontarget for the current image) may require a combination of a number of the approaches above. Future studies will examine how to improve our confidence estimate by combining different approaches from those listed. Such endeavors may provide a more-robust estimate of confidence that will likely help further improve performance.

Once the low performing trials have been identified, one can employ a number of mitigation strategies. The simplest mitigation strategy would be to simply manually label the low-confidence images. If we use the current data to simulate performance when the lowest 20% of trials are manually labeled, overall target-detection error

is reduced by 36%. While the manually relabeling may be the simplest option, it will dramatically increase the time needed to completely label the set of images. For example, Figure 14C shows that approximately 30% of the data must be manually labeled to reduce the nontarget error rate to 20%. If we assume that it takes a user an average of 1 s per manually labeled image, then the manual labeling will increase the total labeling time by 60%. While this increased labeling time may be acceptable for some applications, other strategies may be more efficient. For example, the low-confidence images can be redisplayed to the same person using RSVP, or sent to another person for target identification. Alternatively, we may also be able to couple the human-based target identification with an automatic target recognition system[60,87] to improve performance. Such an endeavor is currently underway[88] and will greatly benefit from the results presented here.

The improvement demonstrated by the inclusion of confidence measures has broad implications for the development of future systems. While we focused on an RSVP-based target-detection paradigm, the use of confidence in human decisions can be extended to a wide range of human-in-the loop systems. The principle of confidence has been applied in control theory to account for variable or noisy sensors. Here we provide initial evidence that the same principle can be applied to account for inherent variability in human decisions.

3.2.3.2 Top-Down Influences

One aspect that was not explored in this study was how top-down influences due to task instructions may have affected performance. In this study, participants were told to explicitly look for people with weapons to test whether the participants and subsequently the classification algorithms could discern people with weapons (targets) from people without weapons (nontargets). The ERP analysis suggests that early stages (200–400 ms) of the P3 waveform may reflect an orienting response to stimulus novelty since rare target and nontarget waveforms were similar but different from the frequent background distractors. Later stages (400–600 ms) of the P3 show differences between targets, nontargets, and background distractors indicating processes related to target selection or nontarget inhibition. Since both targets and nontargets shared many properties (appearing infrequently, people), participants may have adopted a strategy to orient to any rare stimulus. Other research that included a nontarget stimulus in a standard oddball paradigm showed that nontargets have a neural response similar to the frequent background distractors and not the target[89]; however, the stimuli used in this study were simple-shape stimuli containing different stimulus properties (e.g., circles, squares, triangles). This may have lead participants to select targets or possibly inhibit nontarget at an earlier stage of processing than what was seen in the current study.

The ERP waveforms and classification results may have been different if participants searched for targets that did not contain features similar to nontargets[68] or if the instructions had been to explicitly look for weapons (with no mention of people).

3.2.4 Conclusion

By evaluating the impact of adding a nontarget stimulus to a standard RSVP-based paradigm, this study begins the process of moving RSVP based target identification applications into more complex environments that include natural images. We have shown that the introduction of a nontarget stimulus yields a significant slowing of reaction time and reduction of d´. This decrement in behavioral performance is accompanied by a decrement in classification accuracy for single-trial detection and an increase in misclassification rates. Importantly, we show that incorporating measures of confidence can identify trials where the drop in performance is likely to occur. Using confidence measures, we enable these systems to employ a number of possible mitigation strategies that enable the integration of information from alternative sources as a means to improve classification performance.

Testbed Task 2: Exploration

3.3 Leveraging Human Perception to Improve Robotic Exploration and Mapping

The just discussed target detection performance work assumes an incoming stream of images that may or may not include an object of interest. Many such image streams exist, including satellite and aerial photography. Automated, ground-based exploration robots enable more-detailed surveillance of mixed indoor–outdoor environments, but maps of unknown, complex environments may not exist. This introduces complication to target localization/identification because a map needs to be constructed simultaneously. This combination of tasks is referred to in the literature as simultaneous localization and mapping. When imagining a heterogeneous team of automated and human image analysts, some degree of image transmission will be necessary between exploration robots and human analysts. With this infrastructure in place, we sought to determine if human perception could be used to improve mapping as well as target localization.[90]

3.3.1 Background

In autonomous mapping,[91,92] results are dramatically improved when the system is able to successfully detect loop closures when, for example, a previously visited location is recognized as such. This improvement applies to both single- and

multi-agent scenarios. However, recognizing a previously visited location is challenging in general because shifts in perspective can cause the location to appear very differently than it did in the past. For example, autonomous, computer-vision-based systems proposed for this task[93–95] experience difficulty in recognition due to factors such as illumination and viewpoint,[92,96] and overcoming these limitations is still an active area of research. Humans, on the other hand, face and solve the loop-closure problem with what seems to be relative ease throughout the course of normal living. Therefore, it seems natural to wonder whether a joint human–autonomy mapping system can be developed that allows the autonomous system to leverage the human ability to detect loop closures. In fact, exactly such a system was developed as part of Olson et al.[97] with the goal of increasing performance in a map co-registration algorithm. However, this system relied on high-level guidance from expert human users. In contrast, we seek here a technique to gather low-level loop-closure guidance from nonexpert human users.

To do so, we developed a task with which humans could be queried to determine whether or not a robot was visiting a previously visited location. Humans with no special training in, or knowledge of, robotic exploration were shown pairs of video clips recorded by a video camera attached to a ground-based exploration robot taken at 2 different times. The humans were simply asked to rate whether the clips were recorded from the same location or different locations. In this way, we sought to incorporate a task in which we expect most humans to be expert: location recognition.

Given our specific task, the question remains whether human loop-closure skills can be applied reliably when experience of the location is mediated by the limited-field-of-view video recorded by a ground-based exploration robot. Human perception of natural scenes has been studied extensively, and several studies have revealed that human vision can almost immediately understand the gist of a natural scene.[98–100] However, this broad categorization is insufficient for the task of determining whether a particular view or scene matches one in memory. Human scene recognition may depend on encoding the spatial layout of objects in a scene.[101,102] Much of our understanding of human scene recognition comes from research involving highly distinctive objects in tabletop scenes or from static, high-resolution imagery; it is unclear how applicable these findings are to recognizing locations from different views in the real world.[103]

Our video-mediated scene-recognition task differs from naturalistic human location recognition in other ways. For example, location recognition in the real world depends upon recognizing and localizing views from past experience,[104] but matching views in short-term memory, as in our task, likely relies on different

mechanisms.[105] Moreover, in physical navigation, additional cues like path integration might aid with location recognition.[102] In single-robot-mediated exploration, video-mediated visual path integration might be possible but only if the entire putative loop were displayed to the human and other path integration cues like vestibular and somatosensory information would be absent. In multi-robot-mediated exploration, 2 different robots might visit the same location from different directions. In that case, there is potentially no path over which to integrate, and visual recognition might be required to operate in isolation.

To enrich our understanding of how humans might perform our scene-recognition task, we experimentally varied 2 properties of the video clips: clip duration and clip speedup. Duration was varied to see whether additional context would improve scene recognition. Adding context might help if participants used context to build up a mental map of a scene to develop a viewpoint-invariant representation[106–109]; however, participants might alternatively use the presence or absence of specific diagnostic objects or landmarks in a scene to recognize it. In that case, additional context may not help and may instead hurt performance because it adds potentially distracting information in the form of objects and landmarks that might be relatively far away from the target scene. The speedup manipulation was included to help determine if more- or less-dense frame sampling affected performance. By speeding up the clips, human operators may be able to derive the same amount of contextual information in less time and with less image-data transmission.

An additional measure of interest was whether human performance improved over the course of the experiment. We expect that scene recognition of familiar environments should be better than unfamiliar environments, but under the circumstances of ground-robot-mediated exploration it is unknown how quickly that familiarity might form or if such familiarity will help with scene recognition. If ground-robot-mediated exposure to a new environment for tens of minutes allows human observers to learn the features of that environment, we would expect scene-recognition performance to improve over the course of the experiment. However, prior experience of similar environments might transfer, so it could be that there is no room for additional learning in our experiment.

Adding a human in the loop adds considerable overhead (in energy and time) to an exploration and mapping task because communicating relevant information (e.g., images) over long distances requires expenditure of power and because humans are typically slow relative to automated systems. To assess the extent to which that overhead is worthwhile, we compared human performance on the loop-closure task with that of a state-of-the-art, automated, visual loop-closure algorithm (FAB-MAP 2.0).[94,110]

In the following, we present several experimental results for our task of showing pairs of video clips to human observers and asking them to judge whether the clips came from the same location or different locations. We varied clip duration and speedup. We found that humans performed the scene-recognition task well above chance levels and that performance did not vary significantly with clip duration, clip speedup, or learning over the course of the experiment. To assess whether human scene recognition could viably contribute to an automated mapping system, human performance was compared against an automated solution. Human scene-recognition performance was better than automated scene recognition. These results support using human assistance in robotic exploration.

3.3.2 Methods

3.3.2.1 Military Operations in Urban Terrain (MOUT)-Site Database

The video clips used in this experiment were extracted from a data set of robotic sensor readings at a MOUT training site. This data set was captured to simulate an exhaustive exploration of the site. A schematic map and some example views are shown in Fig. 15. Nodes representing a physical location and pose (facing north, south, east, or west) were defined both indoors and out, and nearby nodes, including all at the same location with different poses, were connected to each other by paths. There is approximately one node per indoor room, and exterior nodes were chosen as reasonable waypoints one might visit when traveling between buildings. An iRobot Packbot—equipped with a Prosilica high-resolution color camera (resolution 2752 × 2200; frame rate 1 Hz), an Asus Xtion Pro RGB (red, green, blue)-D camera (resolution 320 × 240; frame rate 30 Hz), an actuated Hokuyo LiDAR (light detection and ranging), and a Garmin GPS—traversed these paths at a speed of approximately 1.2 m/s while recording all sensor data. For the present experiment, only the Asus video images were used.

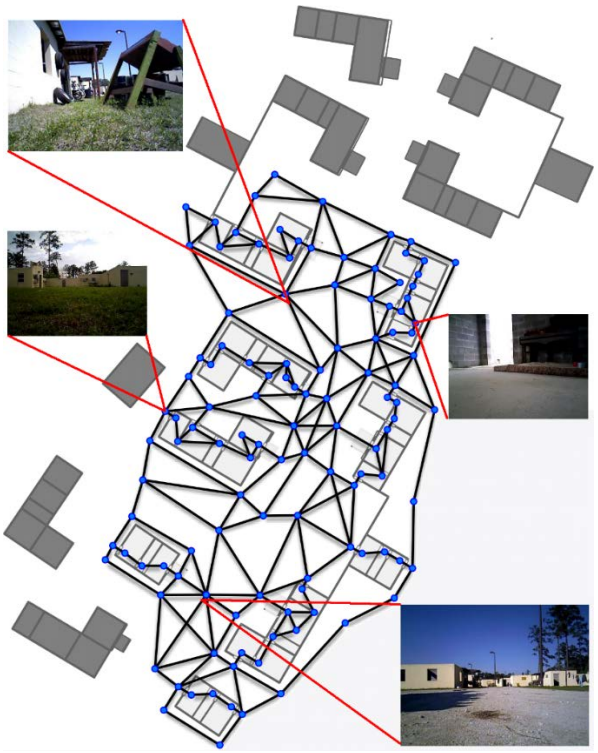

Fig. 15 Schematic and example views from the MOUT database: nodes (blue dots) are both indoor and outdoor locations; connecting paths (black lines) show the exploration robot's paths; selected frames illustrate differences in lighting conditions and scene types.

3.3.2.2 Stimuli, Protocol, and Participants

Trials were presented in blocks by original recording length (2 levels: 6 and 12 s at 30 frames/s) and temporal downsampling (2 levels: 3× and 6×). Temporal downsampling was achieved by only showing every third or sixth frame. Each block consisted of 26 matching pairs and 26 nonmatching pairs presented in random order. Additionally, participants were shown a 3-trial initial block (duration 9 s, speedup 4×; not analyzed) to familiarize them with the experiment interface. The ordering of the blocks was counterbalanced across subjects so that effects of clip types could be examined separately from effects of learning. Clips were displayed in original color at a resolution of 1280 × 960 pixels (upscaled with bilinear filtering from the original 320- × 240-pixel images) at 20 frames/s on an Asus VS248H full high-definition LED 24-inch monitor viewed at 0.7 m. The experiment was controlled in MATLAB using PsychToolBox.[111]

The structure of a trial was the following. The participant pressed a button to indicate readiness, at which point the first clip was displayed. The clip was a segment extracted from the MOUT-site database previously described. After the clip, a black screen was displayed until the participant pressed a button to indicate readiness, and then the second clip was displayed. In matching trials, the second

clip included video recorded while the robot was at the same location (although not necessarily facing in the same direction) as in the first clip. In nonmatching trials, the second clip did not include any video from any node in the first clip. Nonmatching clips only included nodes that were at least 3 paths away from any node in the first clip. After the second clip, the participant was prompted to respond using buttons marked strong nonmatch, weak nonmatch, weak match, or strong match. After the participant responded, he/she was prompted to start the next trial.

The voluntary, fully informed written consent of participants (N = 21) in this research was obtained as required by federal and Army regulations.[112,113] The investigator adhered to Army policies for the protection of human subjects.[113] All human subjects testing was approved by the Institutional Review Board of the US Army Research Laboratory (ARL).

From the participant responses, the area under the Az was calculated for each clip type using Signal Detection Theory (SDT)[114] to fit an Az to the observed responses using the unequal variance normal assumption.[115] With 4 response buttons, 3 criterion observations were possible, so a 2-parameter fit could be made. If a subject did not use all 4 buttons, a reduced ROC model using a unit-variance assumption was fit. Az was calculated as the analytical integral of the fit ROC.

3.3.2.3 Automated Loop-Closure Baseline

FAB-MAP 2.0[93,94] was used as a baseline for automated loop closure as implemented in Glover et al.[110] The scenario presented to the subject in this experiment is different from the typical automated loop-closure-detection scenario in that the automated solution was designed to compare a current view against all previous views. As such, this work should not be seen as an evaluation of FAB-MAP 2.0; rather, it serves as a convenient baseline for comparison against human performance. To adapt FAB-MAP 2.0 to the current scenario, we computed a scene-similarity matrix for each pair of video clips by computing the similarity of each frame from the first clip against all of the frames from the second clip. This produced scores for both matching and nonmatching pairs. The maximum score in the similarity matrix was used as the overall score for the trial. This allowed for the fact that pairs of clips were considered matching if any portion of the clips overlapped in location. The SDT unequal-variance assumption did not fit the obtained ROC (see "Results", Section 3.3.3), so area under the ROC curve was estimated using trapezoid integration.

Just as with the human participants, we also wanted to determine if previous exposure to a particular environment improved the algorithm's performance. We therefore evaluated automated performance in 2 training conditions: generic and

site specific. For the generic condition, the vocabulary was generated from a video gathered during a car ride through an urban environment.[116] For the site-specific condition, the vocabulary was generated from views of the mock urban site that were not used in our experiment. The same number of training frames were used in both training conditions.

To establish a fair baseline for comparison, the automated loop-closure detectors were tested by presenting them with the same clip pairs that human participants saw. This allowed us to extract automated performance for each block, resulting in 21 participants by 4 conditions by 2 training conditions estimates of Az.

3.3.3 Results

3.3.3.1 Characterizing Human Scene-Recognition Performance

Results (summarized as Az) for humans, site-trained automation, and generically trained automation for the 4 video clip conditions are in Fig. 16. The absence of learning over time is illustrated in Fig. 17. Across all stimulation conditions, human performance was high with mean Az = 0.865 (SD = 0.047, range [0.778 0.939]). To assess differences in human performance due to clip parameters, a repeated-measures ANOVA was run with factors of original clip duration (6 s, 12 s), clip speedup (3×, 6×), and block order (first, second, third, fourth). None of these factors had a statistically significant effect on Az (all $p > .05$). There was a statistically reliable effect of subject, $F(20, 57) = 1.86$, $p = 0.035$, $\eta^2 = 0.39$, indicating that between-subject variability was significantly greater than within-subject variability. An identical repeated-measures ANOVA using a logit transform of Az as well as repeated-measures ANOVAs using Az and logit-transformed Az as computed using trapezoid integration were run. All of these analyses had equivalent results.

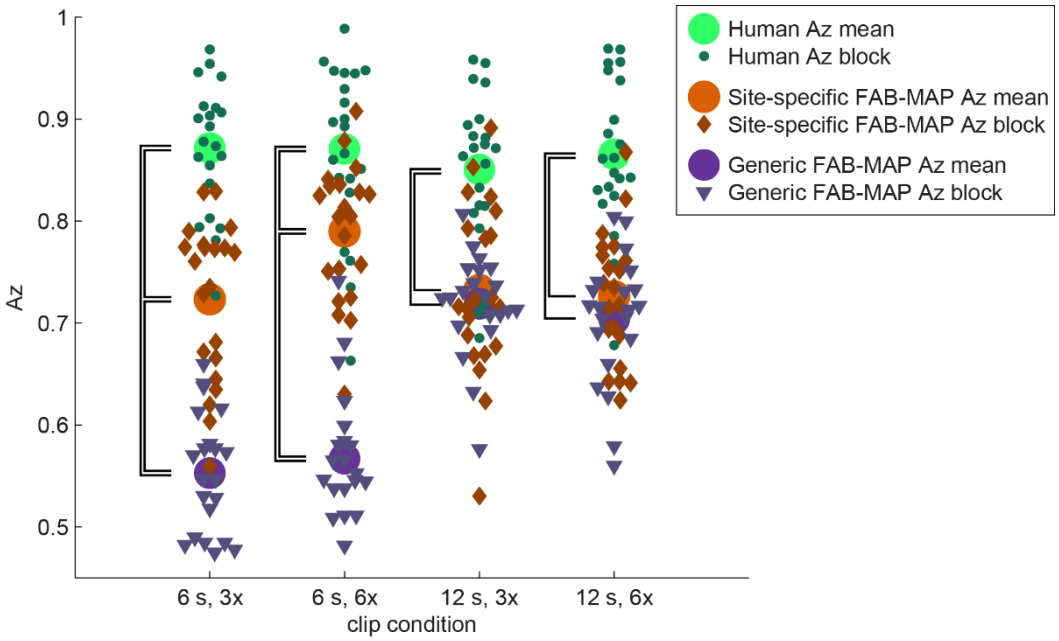

Fig. 16 Human, site-specific (trained), and non-site-specific (generic) automated scene-recognition performance: large dots show means taken over all subject-specific clip sequence blocks; smaller symbols show results for each subject-specific sequence block; brackets indicate statistically significant differences in mean Az as determined by a paired-T test (df = 20, p < 0.01, Bonferroni corrected).

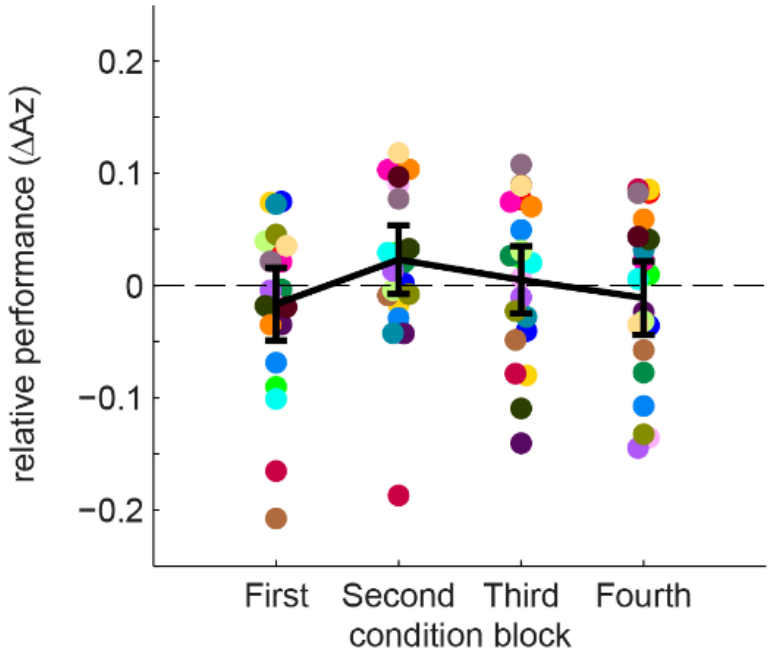

Fig. 17 No statistically significant learning was observed over the course of the experiment. Each dot (one color per participant) shows Az on an experimental block minus the group average performance for that block's video-clip condition. Error bars show approximate 95% confidence intervals for the mean.

3.3.3.2 Automated Scene-Recognition Performance

For the automated scene recognition, the unequal-variance assumption did not result in an ROC-curve estimate that matched well with the empirical ROC (Fig. 18), especially from the 2 shorter clip conditions, so trapezoid integration was used to estimate Az. Mean Az from our adaptation of the FAB-MAP 2.0 algorithm to our task with generic training was 0.636 (SD = 0.034) and with MOUT site-specific training was 0.743 (SD = 0.035). Figure 16 shows results for automated scene recognition in context with human performance. To assess the effects of our experimental manipulations, an ANOVA was run on the automated scene recognition Az estimates with factors of duration, speedup, and training type. An additional factor of clip set was included because even though it was the same algorithm evaluating each subject's particular set of video clips, we wanted to account for the possibility that our pseudo-random sampling of clips resulted in more or less difficult clip sets. Full results are in Table 6. Significant ($p < 0.05$) main effects of training and clip duration were found, but there were also significant interactions of clip duration with speedup and of duration with training type.

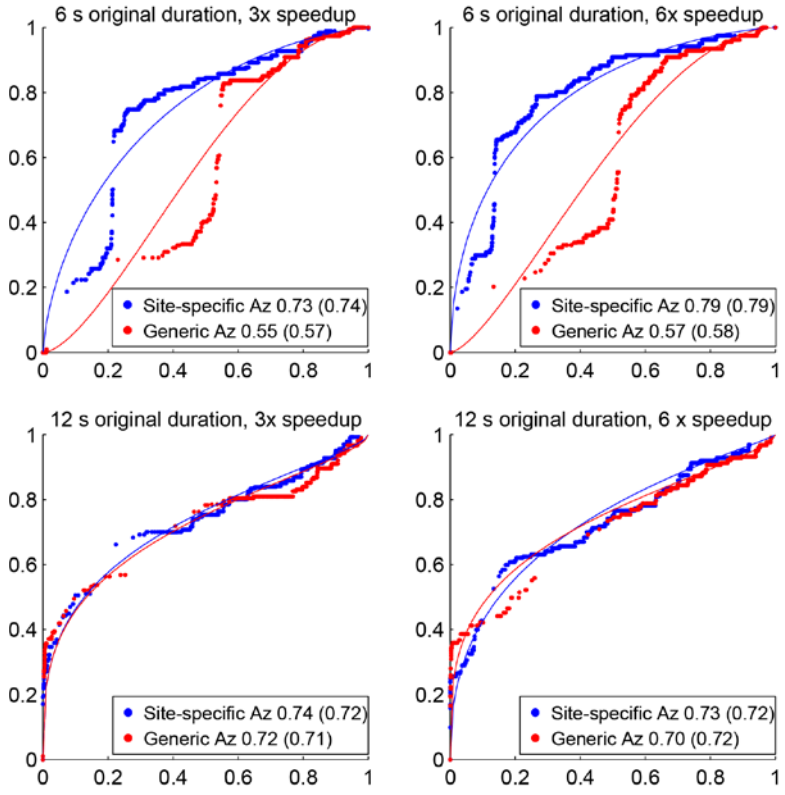

Fig. 18 ROC curves for site-specific trained (site) and site-nonspecific trained (generic) automated scene recognition. Performance shown is over all clip pairs of the indicated length and speedup. Fit lines show the unequal-variance normal distribution curve of best fit. Numbers in the panel legends indicate area under the ROC curve by trapezoid integration and by SDT in parenthesis.

Table 6 ANOVA result for automated scene-recognition performance

Source	Sum sq.	d.f.	F	p-value	eta squared
Clip duration	0.1616	1	38.19	0.000	0.092
Speedup	0.0093	1	2.19	0.141	0.005
Training	0.4831	1	114.17	0.000	0.274
Duration*speedup	0.0273	1	6.46	0.012	0.016
Duration*training	0.3349	1	79.16	0.000	0.190
Speedup*training	0.0099	1	2.35	0.128	0.006
Duration*speedup*training	0.0049	1	1.17	0.282	0.003
Clip set	0.1381	20	1.63	0.053	0.078

3.3.3.3 Man vs. Machine vs. Machine

To statistically assess differences in scene-recognition performance among humans, site-trained FAB-MAP, and generically trained FAB-MAP, we used a family of paired-*T* tests, controlling for differences in clip set difficulty. Within each clip condition, we compared human versus site trained, human versus generically trained, and site trained versus generically trained. To correct for multiple comparisons, we used Bonferroni correction for 12 comparisons (3 contrasts per condition × 4 conditions). In all 4 conditions, human performance was significantly better than both CV instances, and site-trained performance was better than generically trained performance for the two 6-s clip conditions (all $p < 0.01$, Bonferroni corrected). Performance was not significantly different between the 2 CV conditions for the 12-s clips.

3.3.4 Discussion

In summary, we found that human scene recognition was good and did not vary with respect to video-clip length, video speed, or time on task. We also found that human scene recognition was better than automated scene recognition. These results suggest that incorporating human scene recognition into human–autonomy teams for exploration and mapping is a promising way forward. Here, we first consider the results pertaining to the capacity of humans to complete our clip-matching task. We then discuss these findings in the context of human–autonomy teaming for the task of loop closure.

3.3.4.1 Human Scene Recognition

Our results showed that humans are able to accurately recognize scenes from different viewpoints using only brief video clips recorded by exploration robots. The average area under the ROC curve of human sensitivity to the distinction

between 2 views of the same scene versus views of 2 different scenes was 0.865. There was significant inter-individual variability, suggesting that, in a human–autonomy integration application, selecting a high-performing human operator should have even better results. The best performer in our sample had an estimated Az of 0.939. However, this performance is not perfect. Future work will examine whether scene-recognition accuracy in this range is sufficient to improve overall map accuracy.

We found no statistically reliable difference in human scene-recognition performance based on video-clip duration. This suggests that the additional spatial context derived from the longer clips did not provide additional useful information or that whatever useful information they provided was counteracted by distracting information. This result is compatible with an account of scene recognition based on diagnostic objects or landmarks in the target scene. Longer clips may have provided views of more objects or landmarks overall, but objects viewed while approaching a scene from one direction might not be visible when approaching from another direction, so the additional objects might overload working memory without affording any performance improvement. Future work will use eye tracking to infer whether fixation on the same object viewed in a pair of clips is associated with accurate scene recognition. We also found no reliable effect of clip speedup. This indicates that, at least in the limited range we tested, human time efficiency can be boosted by speeding up video recorded by exploration robots.

On average, human performance did not change significantly over the course of the experiment. For this observation there are 2 primary explanations. The first possibility is that the natural statistics of the scenes in the experiment were sufficiently similar to those in the human observer's experience that there was effectively nothing for the observer to learn. The other possibility is that there is something to be gained from extensive experience of a specific environment but that learning takes place at a different time scale, possibly hours or days.

In the loop-closure literature, loop-closure performance is often given as recall at 100% precision. This is equivalent to hit rate at FAR = 0. Our model of human performance from SDT prohibits FARs of 0, so this measure cannot be readily computed. However, humans typically were able to use our 4 response buttons appropriately; increasing certainty of match corresponded to an increase of precision and decrease in recall. Perhaps precision of 100% could be obtained by giving humans a button marked "100% definitely a match".

To summarize the human scene-recognition results, human performance was approximately equally good over the conditions we tested. Although increasing the number of humans in a system is not as easy as increasing the number of robots,

the task humans performed here requires essentially no training other than that obtained in the course of normal daily life. Although human performance was fairly high, it is important to know if human performance was better than performance by automated solutions.

3.3.4.2 Comparison with Automated Scene Recognition

To address the question of whether human input generated by our task might be useful in automated scene recognition, we compared human performance with a state-of-the-art, vision-only scene-recognition algorithm: FAB-MAP 2.0. In an attempt at a fair comparison, 2 variants of this algorithm were used. The first used a visual vocabulary that was generated by analyzing a generic video clip. The second variant used a visual vocabulary generated from analyzing images of our specific experiment environment. Neither variant resulted in performance that approached that of our human participants. However, FAB-MAP 2.0 was not specifically designed for the task here, so additional development might improve automated solutions. Also, in a practical application, an exploration robot would potentially have access to odometry, pose estimation, and nonphotographic measurements of the environment (e.g., LiDAR). These additional information sources would presumably improve automated scene recognition to some extent.

We examined automated scene recognition with generic and with site-specific training, and we found that overall the best automated performance was obtained with site-specific training. But the difference between site-specific and generic training was largest for the shorter, more-sped-up clips. On the longer clips, the difference between specific and generic training was negligible. It was somewhat surprising that the generic training was able to perform as well as it did; the generic training video was recorded by a car travelling on paved roads through a suburban environment, while the testing video was recorded by a slow-moving robot travelling through both indoor and outdoor scenes. This result suggests that training on a large set of images from varied environments might be suitable for exploration in novel environments, at least under some conditions. However, site-specific training outperformed generic training on the shorter video clips, and automated performance was best for the site-specific training on the shortest, most-sped-up clips.

One explanation for obtaining the best performance on the clips with the fewest frames available is that any pair of randomly selected frames has a chance of returning a strong, false match. This is referred to as perceptual aliasing. Our clip-matching score was computed as the peak frame-matching score. This strategy could be unnecessarily sensitive to FPs, and a method that models trajectories, for example,[117] might reduce FPs. However, the specific task in this experiment is

about identifying when clips show the same location and is explicitly not about whether the clips follow the same trajectory, so trajectory matching might also reduce the hit rate of an automated solution for this specific problem. It was not surprising that site-specific training improved performance, but it was somewhat surprising that the advantage of the site-trained algorithm vanishes for the longer video clips. It might be that with longer clips, the perceptual aliasing hazard washes out any advantage of site-specific training.

3.3.5 Summary

Fully autonomous robotic operation allows for efficient scaling of robot team size[97]; however, fully autonomous and highly accurate performance on perceptual tasks such as object and scene recognition has not yet been achieved, especially with unconstrained task parameters. The upper limits of performance on such tasks are unknown, but human performance on these tasks serves as an example of performance that is better than current autonomous solutions. This observation motivates at least 2 lines of research. The first is toward improvement of autonomous solutions; a reasonable goal is to match human performance for the task we used here. Second, in parallel to improvements in autonomous solutions, research is needed to identify when and how to efficiently include human perceptual decisions in human–autonomy teams.

In conclusion, we developed a task that required no special expertise beyond that typical of human vision to aid in autonomous robotic mapping. We showed that human performance on this task was high and did not significantly vary with changes in video-clip duration, speedup, or learning. We compared performance with state-of-the-art autonomous scene recognition and found that human performance was better. More work is needed to determine the extent of improvements human vision can offer over strictly autonomous solutions, but based on these results, incorporating human scene recognition is a promising approach for human–autonomy teams.

Moving forward, this experiment was replicated with an additional 14 participants from whom we simultaneously recorded EEG and eye-tracking data while they performed the loop-closure task. The EEG and eye-tracking data will be analyzed to assess whether these features can be used to construct a confidence metric on human loop-closure performance. An initial step in this direction was undertaken by an undergraduate summer research intern supervised by HIVE team members. The intern found that fusing pupil-size measurement with behavioral responses resulted in a small but statistically significant improvement in loop-closure classification.

4. Sensor Fusion/Computer Vision

A critical barrier for fielding autonomous systems is the issue of human–autonomy integration. Effective methods for fusing information from multiple disparate sensor modalities are limited. Of specific interest here are methods that enable appropriate fusion of inputs from human and autonomous systems to enable effective leveraging of the specific strengths of each agent. We focused on developing novel fusion techniques that can dynamically incorporate information from a changing environment and changing human performance in an effort to rapidly adapt to these changes in real time. We chose to start with CV as the exemplar autonomous technology as a complement to the human target-identification studies carried out under the human variability section. Recently, CV algorithms have dramatically improved enabling an unprecedented level of accuracy in understanding the contents of images. Nevertheless, most algorithms are unable to function in highly dynamic and cluttered environments. In the following sections, we first describe a novel fusion method known as Dynamic Belief Fusion (DBF).[118] Then, we demonstrate that applying this novel method to combine human and CV inputs can significantly improve target-detection performance.[119] Finally, we show that this new method can also be applied across a range of tasks for improving performance.[120]

4.1 Dynamic Belief Fusion

Current methods for fusing multiple object detectors are often specific to a subset of detectors with shared features.[121–124] However, the field of object detection is undergoing a state of rapid advancement.[125–127] Many detection algorithms and, hence, feature-specific fusion algorithms, are quickly becoming obsolete. There is an increasing need for fusion methods that can combine object detection algorithms regardless of their structure. One effective solution in this case can be late fusion, a process which conditions the "trust" in individual detector outputs on their prior performance, and then intelligently combines the trust-weighted outputs.

Several approaches to late fusion exist, including Bayesian fusion and Dempster–Shafer Theory (DST) fusion. However, Bayesian fusion typically does not yield significant improvements in performance due to its inherent characteristics. Bayesian fusion handles uncertainty in a detector's output by associating a probability to each hypothesis (e.g., 30% chance of target and 70% chance of nontarget); however, the Bayesian approach does not indicate the level of trust to be placed in the probability assignments themselves. Belief theory, a component of DST developed by Shafer,[128] takes a step in the right direction to address the ambiguity in detector quality through its use of compound hypotheses. In

considering 2 hypotheses, target and nontarget, Shafer's belief theory assigns probability to the information that directly supports the target and nontarget hypotheses, and also instantiates an intermediate state, target or nontarget, with its own probability quantifying the level of ambiguity that makes either hypothesis plausible. In this manner, a detector output with a high level of ambiguity can be ignored/down-weighted in favor of a more trustworthy, low-ambiguity detector output. However, assigning these belief probabilities is not a trivial task, and choice of assignment method is critical to fusion performance.

We propose a novel approach, DBF, which assigns probability to hypotheses dynamically under the framework of DST. In this approach, trust in an information source is characterized as a continuous function of its output by assigning a corresponding set of probabilities to each output value. The DBF process is partly illustrated in Fig. 19, in which 3 heterogeneous detectors generate scores for a target-candidate window. Similar to other late fusion methods, these scores are cross-referenced with the detectors' trust models to obtain a set of probability assignments, essentially reweighting the outputs of each detector. The probabilities from multiple detectors are then combined into a single fused detection score via Dempster's combination rule.[129]

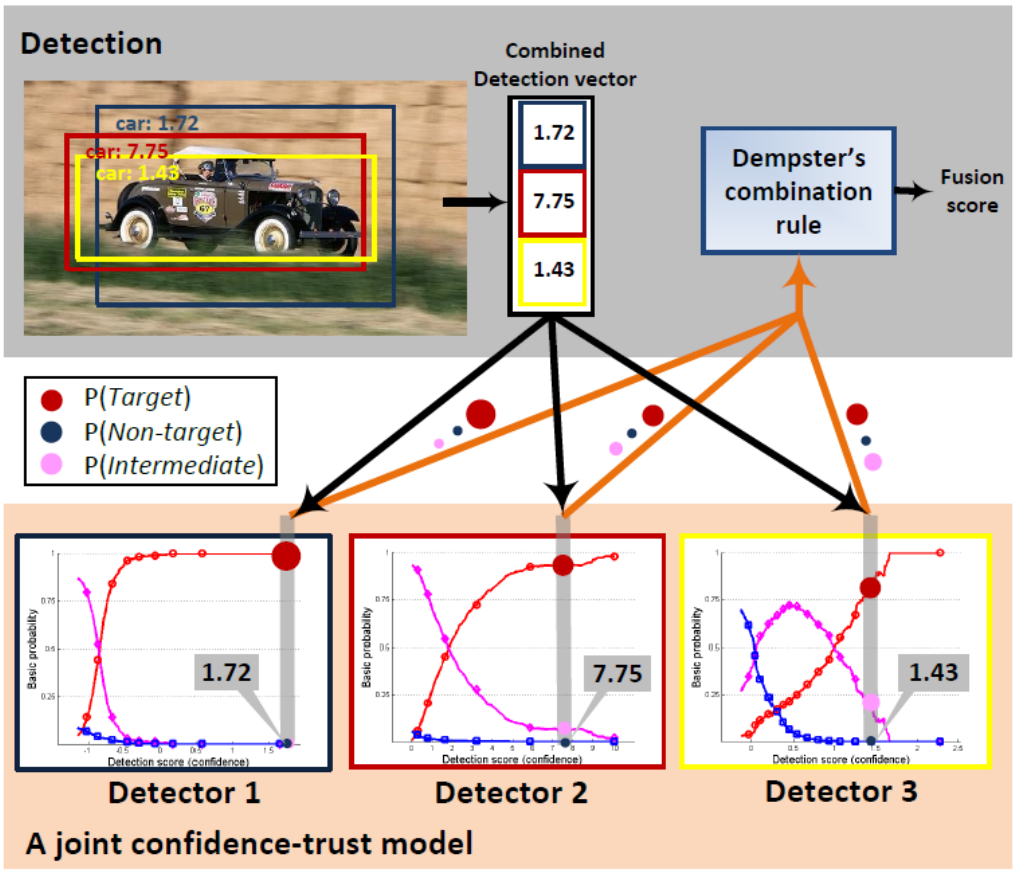

Fig. 19 Dynamic Belief Fusion: 3 detectors—blue, red, and yellow—detect a car in an image shown. A combined detection vector is constructed by collecting detection scores whose windows overlap. For each detector, basic probabilities of target (red), nontarget (blue), and intermediate state (target or nontarget—pink), shown at the bottom, which dynamically vary as a function of detection score in conjunction with the trust model representing prior information of each detector, are assigned. (In each plot, the circle radius represents magnitude of basic probability assignment.) Dempster's combination rule combines basic probabilities of each detector and returns a fused confidence score.

To generate continuous probability assignments (also known as belief functions or trust models) for target, nontarget, and intermediate-state hypotheses in the context of object detection, we employ the precision recall (PR) model of each detector in a validation step. Specifically, to compute the probability assigned to the intermediate state, we devise the notion of a best possible detector, a theoretical detector trained over a limited number of images that can generate the best detection performance possible close to a theoretical limit. We estimate the PR curve of the best possible detector and treat the difference in precision between any individual detector and the best possible detector as the ambiguity in the decision of the individual detector and, thus, assign it to the probability of the individual detector's intermediate state.

Our contributions are summarized as the following:

- We introduce a novel late-fusion framework by optimally modeling joint relationships between a priori and current information of individual detectors. The proposed fusion approach can robustly extract complementary information from multiple disparate detection approaches consistently generating superior performance over the best individual detector. We believe these results, as well as the clear improvement over existing late-fusion algorithms, will inspire greater efforts along the lines of late-fusion research.
- Our novel approach computes the probabilities dynamically, which are assigned to all constituent hypotheses, including an intermediate state (target or nontarget), by optimally linking the current confidence levels in detection (i.e., detection scores) to the PR relationships estimated from a validation set as prior information.

The proposed DBF method is evaluated using ARL[130] and PASCAL VOC 07[131] image sets. DBF is compared with other well-known fusion methods. In these experiments, DBF outperforms all individual CV methods as well as other fusion methods.

4.1.1 Related Works

We can split the literature concerning the fusion of multiple heterogeneous information sources into the following 2 categories:

- building a joint model by integrating multiple approaches
- fusing the output of multiple approaches

Kwon and Lee proposed 2 approaches integrating multiple sample-based tracking approaches using an interactive Markov Chain Monte Carlo framework[132] and using sampling in tracker space modeled by Markov Chain Monte Carlo method,[133] respectively. Wu et al.[134] introduced an approach combining detectors of different modalities (concept, text, and speech) by using relationships among the modes in the event detection. However, in general, modeling the dependencies in fusion among multiple approaches built on different principles is infeasible.

In the case where modeling dependency among multiple approaches is not possible, fusion can be performed over their outputs (late fusion). Bailer et al.[135] introduced a fusion framework in the target-tracking paradigm. They collected trajectories from multiple tracking algorithms and computed one fused trajectory to improve accuracy, trajectory continuity, and smoothness. However, since temporal

information obtained from trajectory cannot be applicable in the object-detection task, a different fusion framework is necessary for our problem. Kim et al.[136] and Liu et al.[137] used weighted sum (WS) methods to fuse multiple types of data for object detection. Their WS method learns weights in a manner that estimates trust in multiple data sources. However, since weights optimization is usually performed to maximize distance between positive and negative samples, like Bayesian fusion, WS does not provide a way to indicate ambiguity between the positives and negatives, which degrades fusion performance, as previously mentioned. The works of Ma and Yuen[138] and Liu et al.[139] employing Bayesian fusion also show limited performance.

To improve upon these late-fusion results, we introduce DBF, a general fusion framework for object detection, employing DST to interpret and leverage ambiguity more completely. Experiments demonstrate the prominent performance of our proposed approach against WS and Bayesian fusion, as well as other existing methods.

4.1.2 Dempster–Shafer Theory

In this section we detail components of DST, which form the foundation of our proposed DBF method. DST[128,129] is based on Shafer's belief theory[128] that obtains a degree of belief for a hypothesis by combining evidences from probabilities of related hypotheses. DST combines such beliefs from multiple independent sources using a method developed by Arthur Dempster.

4.1.2.1 Shafer's Belief Theory

Let X be a universal set consisting of M exhaustive and mutually exclusive hypotheses ($X = \{1, 2, \cdots, M\}$). The power set 2^X is the set of all subsets of X. Basic probability in the range [0 1] is assigned to each element of the power set 2^X. A function defined as m: $2^X \rightarrow$ [0 1] is called a basic probability assignment (BPA). Subsets consisting of compound hypotheses in X represent ambiguity among the constituent hypotheses; the BPA given to the subset measures the level of ambiguity. A BPA has 2 properties: (i) $m(\emptyset) = 0$ (the mass of the empty set is zero) and (ii) $\sum_{A \in 2^X} m(A) = 1$ (the BPA values of the members of the power set sum to one).

From the BPAs, the belief function *bel*(A) for a set A can be defined as the sum of all masses, which are subsets of the set of interest:

$$bel(A) = \sum_{B|B \in A} m(B). \quad (13)$$

Belief represents the information in direct support of A.

4.1.2.2 Dempster's Combination Rule

Dempster's combination rule can be applied to calculate a "joint BPA" from separate BPAs. Under the condition that the evidence from each pair is independent of the other, Dempster's combination rule defines a joint BPA $m_f = m_1 \oplus m_2$, which represents the combined effect of m_1 and m_2:

$$m_f(A) = m_1 \oplus m_2 = {}^{1}\!/_{N} \sum_{X \cap Y = A, A \neq \emptyset} m_1(X) m_2(Y), \qquad (14)$$

where $\mathrm{N} = \sum_{X \cap Y \neq \emptyset} m_1(X) m_2(Y)$ and X and Y are subsets of 2^X. N is a measure of the amount of any mass whose common evidence is not the null set. Dempster's rule can be extended for multiple pieces of evidence (e.g., multiple detectors) using the associative and commutative properties of BPAs, $m_f = m_1 \oplus m_2 \oplus \cdots \oplus m_K$, with the following formula:

$$m_f(A) = {}^{1}\!/_{N} \sum_{X_1 \cap X_2 \cap \ldots \cap X_K} \prod_{i=1}^{K} m_i(X_i). \qquad (15)$$

4.1.3 The Proposed Fusion Approach: Overview of the Fusion of Detectors

The proposed fusion of object detectors is performed in 3 steps: 1) individual detectors are trained on the training set, 2) a PR relationship used as a prior information for the fusion is calculated for each detector on the validation set with detection scores and ground truth information, and 3) in testing, each detection score is converted to probabilities associated with corresponding detection hypotheses for all individual detectors. The probabilities are, for each detector, estimated by adaptively linking corresponding detection score to the PR models previously calculated on the validation set. Joint exploitation of the detection scores in the test set and the PR model in the validation set is used to estimate trustworthiness of the detection in conjunction with the general performance of individual detectors. The estimated probabilities for individual hypotheses are separately fused over different detection approaches. Figure 20 illustrates the fusion process of the proposed DBF algorithm. Details of the proposed fusion process follow.

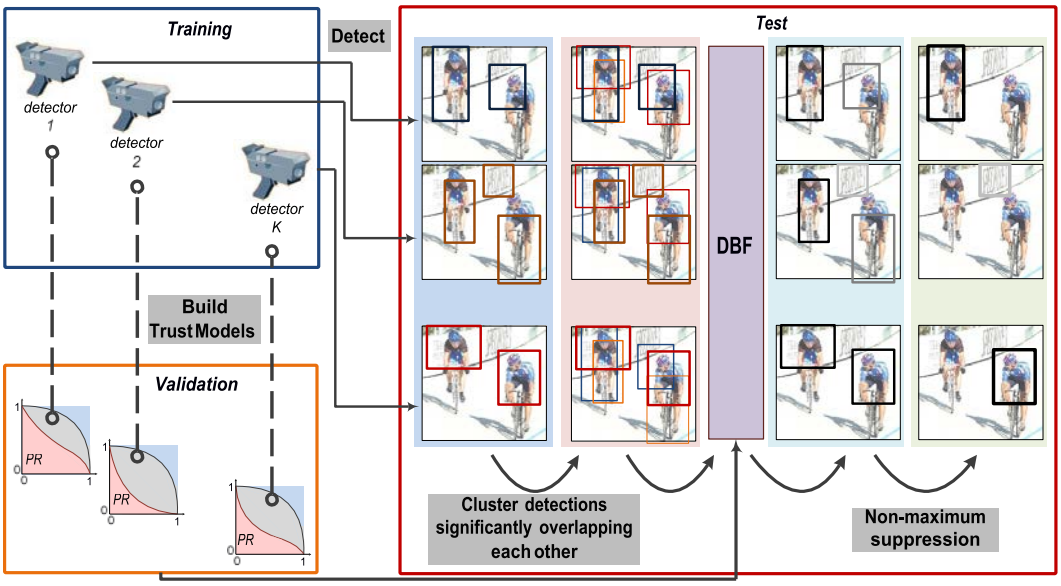

Fig. 20 Flow diagram of proposed fusion algorithm: in fourth and fifth columns (right side of "DBF") of the "test" step, darker windows indicate higher confidence.

4.1.3.1 Building a Trust Model as A Prior Performance Model for Individual Detectors (Validation)

Detectors are applied to validation images in a scanning window fashion and to search for potential objects of interest. To construct the trust model of each detector we first estimate the PR relationships of all the detectors. In building the PR model, all detection windows are labeled as true, FP, or undecided by comparing them with ground truth (annotated windows containing the objects of interest). Any detection window that has an intersection-over-union overlap (PASCAL VOC criteria)[131] of greater than 0.5 with a ground-truth window is assigned true positive. If there is no overlap between a detection window and a ground-truth window, the detection is assigned FP. The remaining detections are labeled undecided. The PR model is constructed using the labeled detection results.

4.1.3.2 Constructing A Combined Detection Vector from Detection Windows for Fusion (Test)

Let $d^i, i = 1, 2, \cdots, K, j = 1, 2, \cdots j, W_i$ be the jth window of the ith detector associated with detection score c^i. K is the number of detectors and W_i is the number of detection windows of the ith detector. For each detection from all of the detectors given a test image we collect the detection windows from the remaining detectors that significantly overlap the subject detection window (see the second column of the test phase in Fig. 20). Two detections, d^i and d^k, $i \neq k$, are considered significantly overlapping if the intersection over-union overlap

of their windows is greater than 0.5. A K-dimensional detection vector c = $[c^1_{1j}\ c^2_{2j}\ ,,,\ c^K_{Kj}]$ is then constructed consisting of the score of the subject detection window and those of the overlapped windows from other detectors. If multiple windows from the same detector overlap the subject detection window, the window with the maximum detection score between them is used. If no overlaps exist for a particular detector, the corresponding element of the combined detection vector is filled by a value of negative infinity to ignore the influence of the detector in fusion.

4.1.3.3 Fusing Detection Windows (Test)

Fusion is performed over the combined detection vector using DBF. Details of DBF are described in the following subsection. DBF dynamically assigns basic probabilities to the hypotheses of a given observation by adaptively mapping current detection scores to the PR model, which are fused over all the detectors by the Dempster's combination rule. After rescoring all windows by applying DBF, nonmaximum suppression is applied to merge windows whose intersection over union overlap is greater than 0.5. The final output of the fusion procedure is a consolidated set of windows, each with a fused detection score.

4.1.4 Dynamic Belief Fusion

In binary object detection, the universal set X is defined as $\{T, \neg T\}$, and thus its power set is expressed as $\{\emptyset, T, \neg T, \{T, \neg T\}\}$, where T is a target hypothesis and $\neg T = X - T$ is a nontarget hypothesis. The $\{T, \neg T\}$ in the power set represents detection ambiguity, denoted by I (intermediate state), which indicates that the subject observation could be either target or nontarget. We assign basic probabilities to all hypotheses based on prior detection performance of detectors. We employ the PR model to represent the prior information of individual detectors and compute basic probabilities of the hypotheses for a given observation. Since the PR relationship is obtained by varying a threshold against detection scores, c^i_j , the basic probabilities being assigned dynamically change as c^i_j changes. Hence, we refer to this assignment as dynamic basic probability assignments.

In DBF, as shown in Fig. 21, each element of the combined detection vector, c^i_j , is first mapped to the corresponding recall and the corresponding precision (p) is assigned as the basic probability of target hypothesis. Then, 1 - p needs to be split to account for 2 basic probabilities of nontarget and intermediate state since it includes information about both hypotheses. This is because precision is only defined for targets (not backgrounds). Note that the recall of background (i.e., recall when "positive" refers to background) cannot be calculated because the number of backgrounds is close to infinite.

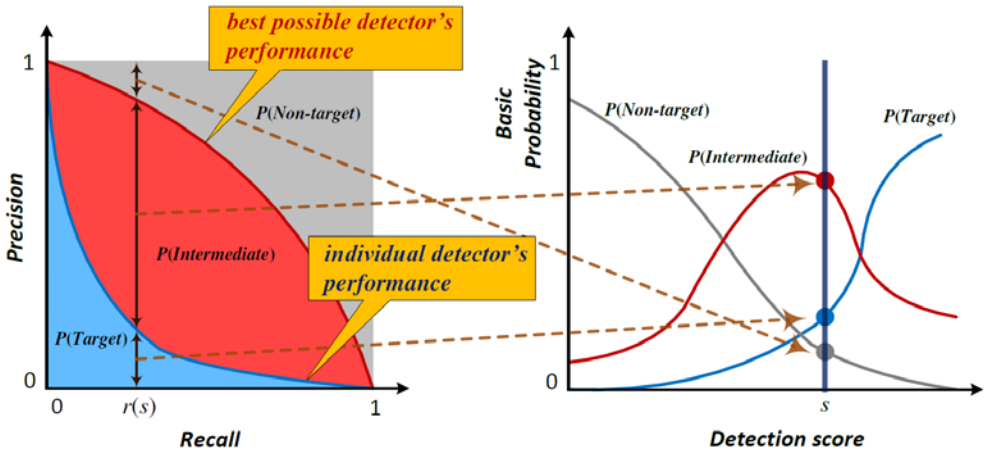

Fig. 21 DBP assignment: Left plot shows a PR curve for an individual detector and a best possible detector. The rates of values along the precision axis corresponding to recall r(s) are assigned as the basic probabilities to target, nontarget, and intermediate state, where s is a detection score. Right plot presents basic probabilities with respect to a detection score, which converted from the PR curve.

Since the split cannot be achieved based solely on the given PR relationship, we introduce a theoretical best possible detector whose performance can possibly achieve a level close to a theoretical limit. Ideally, individual detectors can also achieve the same performance of the theoretical detector if they are provided with complete information about target and nontarget. In reality, individual detectors do not have complete information in training. We treat the difference between the precision of an individual detector and that of the best possible detector as the detection ambiguity (i.e., the probability of the intermediate state) caused by the lack of complete information in training. In our work, the PR curve of the best possible detector, $\hat{p}_{bpd}$, is modeled as

$$\hat{p}_{bpd} = 1 - r^n, \tag{16}$$

where r is recall. This model is proposed because in general $\hat{p}_{bpd}$ should mimic the typical behavior of a highly accurate detector, a concave function approaching the top-right corner of the plot such as the car detector in Dalal and Triggs.[126] The m(I) is defined by $\hat{p}_{bpd} - p$ and the remaining fraction of precision 1 -$\hat{p}_{bpd}$ is assigned to $m(\neg T)$. As n approaches infinity, the best-possible detector becomes the perfect detector (i.e., no FPs). Dynamic basic-probability assignment is shown in Fig. 21. Fusion of the detections from multiple individual detectors is achieved by computing fused basic-probability assignments of target and nontarget hypotheses, $m_f(T)$ and $m_f(\neg T)$, by Dempster's combination rule in Eq. 15. The overall fusion score is given by $s = bel(T) - bel(\neg T)$, where in our experiments, $bel(T)$ and $bel(\neg T)$ are actually $m_f(T)$ and $m_f(\neg T)$, respectively, according to Eq. 13 since T and $\neg T$ are sets of a single element.

4.1.5 Experiments

4.1.5.1 Evaluation Setting

Eight object detectors with unique detection structures whose codes are readily available online were selected: 2 support vector machine (SVM)-based detectors incorporating both histogram of oriented gradient (HOG)[126] and dense scale-invariant feature transform (DSIFT),[139] 2 deformable part models (DPMs) with HOG [127] and color attribute,[140] transductive annotation by graph (TAG),[141] exemplar SVM,[125] and 2 convolutional neural network (CNN)-based detectors (fine-tuned CNN[142] and regional [R]CNN).[143] Given an image, the detection score indicates a degree of confidence about the decision. The 8 selected detectors use different feature-extraction methods (e.g., HOG, DSIFT, color attributes, and CNN features) and different principles of detecting objects of interest.

Baselines (Fusion)

As a baseline, we used 5 approaches: Platt scaling,[86] WS, Bayesian fusion, local expert forest (LEF),[144] and Detect2Rank (D2R).[145] The Platt scaling learns a logistic regression model on the detection scores of true and FP detections. We applied Platt scaling to all the detectors on validation images. At test time, detections from multiple different detectors can be reconciled by fitting the distribution of detection scores of each detector to that of the Platt-scaled validation set. After scaling, the maximum value of the combined detector vector c is used as the final fused score. The WS approach finds weights of detection scores that maximize the product of a weight vector w and the detection vector of detector scores $_c$, $f_{WS}(c) = w^T c$; w is learned through linear SVM optimization. In WS, detection scores are converted into probabilities by Platt scaling as well because negative infinity scores in the combined detection score for the non-overlapping windows can hurt the SVM optimization. For Bayesian fusion, we use a naive Bayesian model assuming that all the approaches are independent of each other. In other words, the joint likelihood can be decomposed as the product of the likelihoods of each detector, while the posterior is expressed as the product of the prior and the joint likelihood (i.e., Bayes' rule). The remaining 2 approaches, LEF[144] and D2R,[145] have been recently introduced. Karaoglu et al.[145] implemented 4 ranking approaches, and we have used PoW2, the best among the 4. These current works are compared with our proposed algorithm only using PASCAL VOC07 data set.

To demonstrate the advantages of dynamic basic probability assignment in the proposed DBF, we also implement a regular DST fusion method that employs only static basic probability assignment,[146] in which each detector's previous

performance is characterized by the probabilities of the 3 hypotheses at the fixed precision value corresponding to a recall of 0.2.

4.1.5.2 Evaluation of ARL Dataset

The ARL image data set was originally created for the purpose of analyzing human performance in RSVP[130] tasks, but is also applicable to object-detection tasks. (In future work, we plan to integrate computer-vision-based object detection with human decisions.) The dataset contains 3000 images of both indoor and outdoor scenes, 1438 images of which contain at least one object of interest. The target objects include chair, container, door, poster, and stair. Figure 22 displays several example images of all 5 objects as well as background images. The number of images in the ARL dataset is relatively small compared with that of other benchmark datasets such as PASCAL VOC 07 and ImageNet. However, with regard to the mean average precision (mAP), the ARL dataset (0.253 for DPM) is not considerably less challenging than the benchmark datasets (0.239 for DPM on PASCAL VOC 07).

Fig. 22 ARL dataset[130]

The proposed DBF algorithm was evaluated on the ARL dataset and its average precision (AP) was compared to that of 4 individual detection algorithms (the HOG–SVM detector was not used on the ARL dataset) and 4 other "baseline" fusion methods (Platt, Bayes, WS, and DST) for each object class. Results are shown in Table 7.

Table 7 AP on the ARL dataset[27]

	chair	contr	door	postr	stair	mAP
DSIFT	.143	.037	.073	.143	.061	.091
TAG	.045	.128	.165	.066	.008	.082
ESVM	.125	.318	.150	.236	.122	.190
DPM	.188	.396	.194	.342	.143	.253
Platt	.191	.364	.204	.307	.125	.238
WS	.192	.388	.267	.318	.096	.252
Bayes	.244	.424	.281	.341	.089	.276
DST	.234	.314	.230	.247	**.168**	.238
DBF	**.329**	**.451**	**.298**	**.390**	.159	**.325**

4.1.5.3 Evaluation of PASCAL VOC 07 Dataset

The fusion and individual detection methods were also evaluated on the PASCAL VOC 07 dataset.[131] PASCAL VOC 07 provides train, val, trainval, and test, where the trainval set consists of images of train and val sets. While previous works that used individual detectors employed in our fusion method use the trainval set, we learn the detectors and trust models on train and val set, respectively. This split is made to avoid building trust models that overfit the training dataset. Therefore, the performance of the individual detectors used in our work is worse than the performance reported in the original literature with regard to the individual detectors, as we are using a smaller training data set.

The mAP of each individual detector and fusion method is reported in Table 8. To evaluate fusion on the PASCAL VOC 07 dataset, 8 individual detectors (DSIFT–SVM, HOG–SVM, TAG, Exemplar SVM, 2 DPMs employing HOG and color attributes, separately, fine-tuned CNN (FTCNN), and RCNN) were selected and fusion of their detection results was conducted.

Table 8 AP on the PASCAL VOC 07 dataset[4]

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbik	pers	plant	sheep	sofa	train	tv	mAP
HOG	.036	.060	.001	.001	.005	.005	.094	.001	.001	.092	.001	.002	.002	.005	.001	.001	.003	.001	.013	.103	.021
TAG	.019	.051	.009	.002	.002	.028	.022	.080	.002	.006	.056	.032	.020	.085	.051	.002	.001	.010	.020	.014	.026
DSIFT	.081	.024	.017	.004	.002	.080	.118	.142	.005	.097	.109	.128	.040	.037	.076	.002	.059	.102	.122	.028	.064
ESVM	.164	.418	.041	.096	.107	.341	.336	.095	.100	.129	.097	.013	.362	.322	.170	.033	.170	.102	.287	.263	.182
Color attributes	.0.201	.518	.026	.102	.167	.344	.363	.172	.158	.198	.041	.358	.349	.436	.376	.106	.128	.273	.304	.307	.246
DPM	.231	.500	.036	.099	.162	.388	.451	.153	.120	.172	.129	.106	.463	.375	.346	.109	.109	.144	.353	**.333**	.239
CNN	.010	.080	.035	.031	.001	.048	.030	.074	.011	.040	.039	.063	.099	.078	.035	.022	.022	.018	.046	.034	.041[1]
RCNN	.637	.709	**.506**	**.393**	.300	.639	.721	**.601**	.303	**.585**	.458	**.559**	.631	**.681**	.549	.291	.536	.467	.575	.662	.540
Platt	.596	.695	.470	.383	.314	.627	.708	.566	.295	.542	.398	.529	.595	.640	.508	.278	.503	.439	.537	.605	.511
WS	.576	.692	.486	.370	.326	.601	.706	.526	.315	.533	.450	.511	.658	.628	.538	.273	.502	.466	.577	.594	.516
LEF	.606	.671	.441	.366	.291	.624	.721	.503	.300	.571	.444	.463	.621	.615	.524	.276	.503	.488	.528	.628	.510
D2R	.609	.687	.468	.398	.311	**.665**	**.757**	.552	.326	.587	.449	.493	.660	.636	.528	.289	.511	.502	.550	.654	.531
Bayes	.460	.616	.177	.098	.297	.541	.644	.252	.115	.413	.278	.344	.359	.517	.229	.215	.447	.138	.461	.475	.354
DST	.423	.601	.291	.250	.259	.580	.624	.382	.215	.348	.352	.291	.532	.511	.483	.227	.345	.277	.456	.488	.397
DBF	**.650**	**.720**	.501	.392	**.341**	.658	.729	.576	**.339**	.578	**.477**	.537	**.670**	.664	**.572**	**.315**	**.537**	**.539**	**.590**	**.672**	**.553**

4.1.5.4 Discussion

Both mAP and ROC performance metrics show that DBF outperformed all of the baseline fusion algorithms as well as individual detectors on both ARL and PASCAL VOC 07 datasets. DBF demonstrates the best results for 4 of 5 categories in the ARL dataset and 12 of 20 categories in the PASCAL VOC 07 dataset. DBF is the only fusion approaches that outperforms RCNN on the PASCAL VOC 07 though improvement is small. Only a minor improvement is achieved because the performance of RCNN is much greater than the other detectors. Therefore, we evaluated fusion performance again but without RCNN: the mAP for Platt, 0.268; WS, 0.271; LEF, 0.283; D2R, 0.261; Bayes 0.253; DST, 0.257; and DBF, 0.341. DBF still outperformed all baseline fusion methods and all individual detectors with a significant gain in mAP (0.06 from LEF and 0.10 from color attributes). The clear difference in performance between conventional DST fusion and the proposed DBF demonstrates the strength of dynamic basic probability assignment over the conventional method of static assignment. Likewise, the fact that DBF outperforms Bayesian fusion demonstrates the benefits of adding an intermediate state to the set of hypotheses.

In addition, we analyzed top-ranked FPs on the PASCAL VOC 07 and categorize them into 4 types according to Hoiem et al.[147] in Fig. 23: 1) poor localization (Loc), 2) confusion with similar classes (Sim), 3) confusion with dissimilar object categories (Oth), and 4) confusion with background (BG). Notably, most of FP in CNN performance is from poor localization. We can guess that CNN performed much worse than expectation because coarse-grid scanning windows and aspect ratio of windows fixed as square bring localization error. FPs detected by RCNN

have similar fractions to CNN. Once accurate localization approaches replace coarse-grid sliding windows (employed by CNN) or objectness (employed by RCNN), CNN-based detector may achieve much better performance. The charts also demonstrates that, as compared with RCNN, DBF increases the performance by reducing inaccurately localized FPs.

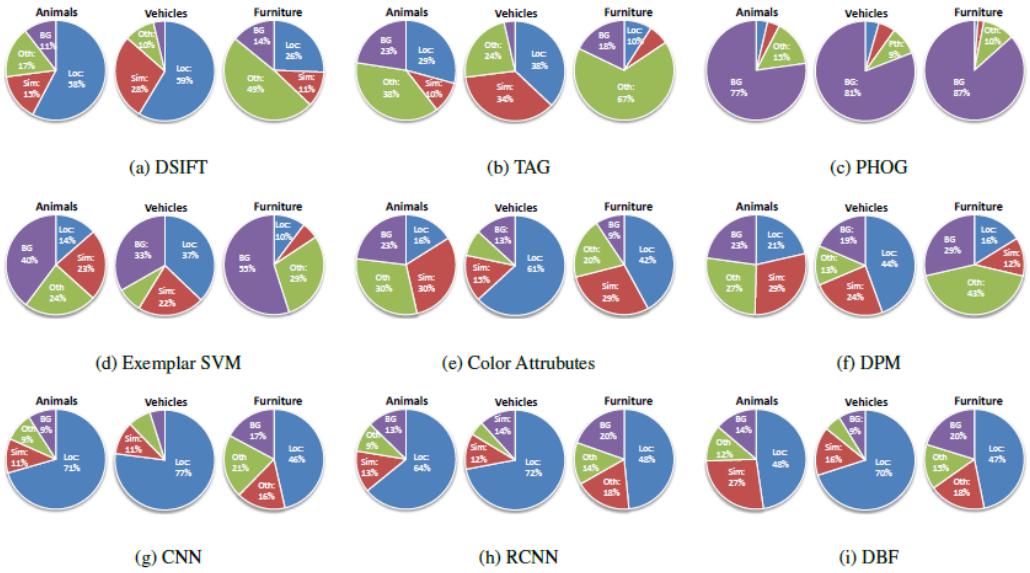

Fig. 23 Analysis of top-ranked FPs: Pie charts present fractions of 4 types of top-ranked FPs. Analysis is performed on PASCAL VOC 07 data set. Among 20 object categories in PASCAL VOC 07 data set, all animals including person are in "Animal"; all vehicles are in "Vehicle"; and "chair", "dining table" and "sofa" are assigned to "Furniture". Loc error, confusion with Sim classes, confusion with Oth categories, and confusion with BG are indicated by blue, red, green, and purple, respectively.

To further investigate whether (and to what degree) complementary information is provided by each detector using DBF, mAP was calculated while varying the number of individual detectors used in fusion. For each combination number K, detectors with the K highest mAP were selected. The results, shown in Table 9 for both the ARL and PASCAL data sets, illustrate that performance improves as the number of detectors increases, at a decreasing rate.

Table 9 Comparison of fusion performance with respect to the combination of multiple detectors

# of detectors	ARL	PASCAL
2	.295	.545
3	.319	.547
4	.325	.548
5		.548
6		.552
7		.553
8		.553

The final row corresponds to the maximum number of combined detectors (4 for ARL, 5 for PASCAL). Figure 24 illustrates the variation in mAP as the shape of the PR curve of the best possible detector is varied for each object category in the ARL dataset. The optimal value of the parameter n (Eq. 4), which dictates the shape (and, hence, estimated performance) of the best possible detector, is different for different object categories. However, the notional perfect detector ($n = \infty$) underperforms other choices of n in every object category. This result suggests that our method of splitting the FPs into nontarget and intermediate state categories is actually beneficial.

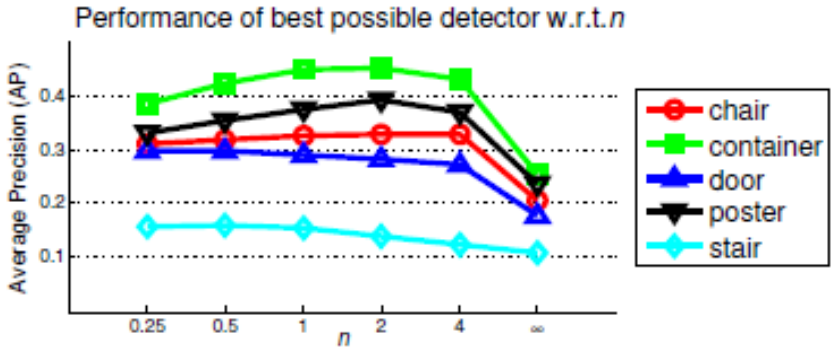

Fig. 24 Comparison of fusion performance with respect to the various theoretical best-possible detectors; n in x axis is the exponent in Eq. 4

4.1.5.5 Conclusions

DBF is proposed to improve upon current late fusion methods in the context of object detection. DBF employs prior information in the form of dynamic basic probability assignments. For object detection, these dynamic basic probability assignments (target, nontarget, and intermediate state) are generated from the precision-recall curve of a validation image set. To properly separate the nontarget and intermediate states, the concept of a best possible detector is introduced and applied. Dempster's combination rule is used to combine the resulting basic probabilities of detections from different detectors.

Experimental results on 2 datasets, ARL and PASCAL VOC 07, demonstrate that DBF outperforms all baseline fusion approaches as well as all individual detectors in terms of mAP DBF also achieved performance improvement over RCNN on PASCAL VOC 07. Its superior performance compared with the DST-based fusion approach (incorporating fixed levels of basic probabilities) clearly illustrates the robustness of dynamic basic probability assignment. Enhanced performance over Bayesian fusion supports the use of an intermediate belief state, which was achieved in this context via the instantiation of a best possible detector.

DBF is a novel approach guaranteed to provide improved fusion performance over the best detector in conjunction with other detectors in the fusion pool through dynamic belief assignments and the Dempster–Shafer combination of assigned probabilities. Therefore, addition and removal of individual detectors from the fusion pool can only further improve fusion performance as state-of-the-art detectors, such as deep learning approaches, are introduced.

4.2 DBF for Joint Human–Computer Vision Image Labelling

4.2.1 Introduction

Human–autonomy sensor fusion combines the raw processing power and consistent response of autonomous systems with the context awareness, vast experience, and adaptability of humans. Of particular interest, considering the current capabilities and limitations of autonomy, is the object detection task, typically defined as indicating the location of a specific target within an image. CV-based methods, which rely on feature extraction and pattern matching, have steadily improved over the past decade[148] and are being used in real-world applications such as facial recognition[149–151] and action/event detection.[152]

Speed and accuracy are both paramount in a wide variety of time-critical object-detection scenarios. While humans possess an unparalleled ability to detect objects in images, the speed at which they are able to report detections is limited. The opposite is true of autonomous systems: With enough computing power, object detection speed is negligible, but accuracy currently does not rival that of humans. Even state-of-the art methods cannot fully account for context and are prone to mistakes in the case of uneven lighting and/or clutter. Most importantly, humans and autonomous detectors are heterogeneous systems that often contain complementary information; therefore, proper combination may significantly improve detection accuracy in comparison to either agent alone.

Similar approaches have been investigated in prior work. Wang et al. performed experiments with a single NC and a graph-based pattern-mining system to identify

relevant images from an image pool, demonstrating improvements over NC alone.[87] Our own previous work used multiple NCs with the same graph-based pattern mining system to show that combining human vision and CV can improve image classification over either source independently.[88] In both of these previous studies, however, the system did not attempt to localize target objects within the images (i.e., detection). Human–autonomy sensor fusion has also been explored in other areas. For example, Fanaee and Gama proposed various architectures for combining data-anomaly detection and background knowledge for event labeling, including the addition of human experts, but did not use images.[153] Prosthetics have been developed that rely on both mechanical and physiological sensor information to better decode user intent.[154–156]

The objective of this work is to assess the benefits of human–autonomy fusion in target detection tasks, as well as the advantages and limitations of various fusion methods when applied for this purpose. We develop a fusion procedure that adapts CV-based target-detection algorithms to incorporate human responses. To capitalize on human detection abilities while maintaining high throughput, RSVP is employed. In most previous cases, RSVP is limited to image classification (target presence or absence) and provides no information about target location. It is important to stress the difference between object detection, which refers to estimating the bounding boxes of each object of a given object class in a test image,[148] and object classification, which only requires predicting the presence/absence of at least one object of a given class in a test image. Our methodology augments CV-based object detection with classification of human response.

To properly fuse these heterogeneous information sources, an intermediate step is introduced in the fusion process whereby detector fusion and classifier fusion are performed separately. Then a final fusion step is implemented to augment detection scores with classifier information, as illustrated in Fig. 25. Human classifiers score on a per image basis, while CV detectors score multiple detection windows within an image. Although CV can be applied to images in real time (e.g., as a human is undergoing RSVP), in this experiment detection was performed offline. Three candidate fusion algorithms were evaluated: 1) Bayesian, 2) Dempster–Shafer, and 3) Dynamic Dempster–Shafer fusion. These "late fusion" methods are independent of the detectors/classifiers employed, and unlike feature-level fusion methods remain applicable as new detection and classification methods are introduced. The fused detection windows generated for each image are compared against ground-truth windows, and PR is employed to evaluate performance.

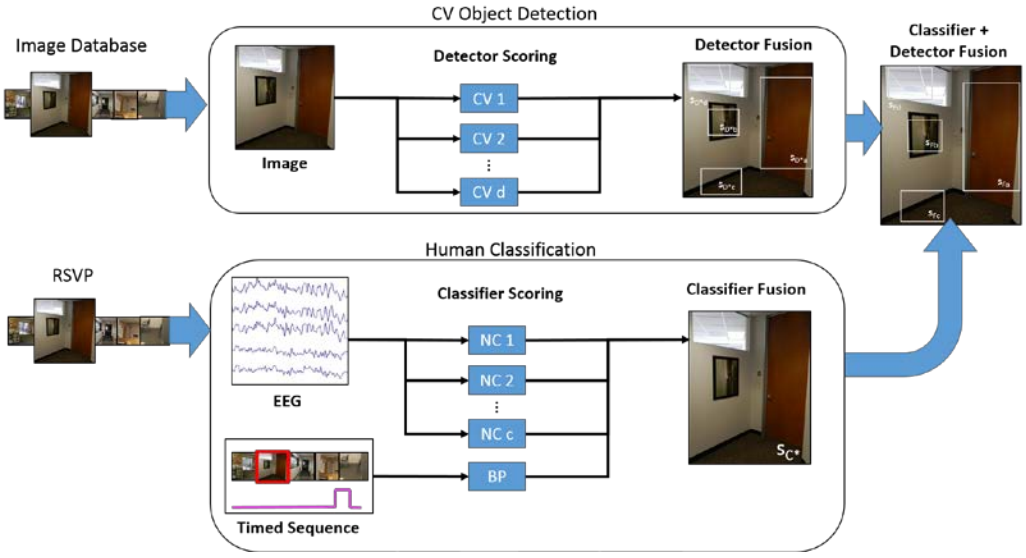

Fig. 25 Fusion of CV-based object detection with human neural and button-press classification; CV = CV detector, NC = NC, and BP = BP classifier

The fusion of human-based classifiers and CV-based detectors generally provide an improvement in average precision over either method alone. This improvement illustrates that our methods can reliably extract complementary human and CV information. In contrast to previous studies, where BP and NCs each contributed to improving image classification,[88] we found that for object detection, BP was shown to provide the greatest benefit, while the XD+BLDA NC also provided similar benefit. Other NCs tested provided negligible benefit and in some cases were detrimental. Results also show an inherent limit to the performance boost provided by classification, indicating that localization information from human eye tracking may be beneficial in future work. A novel method of fusion from Lee et al.,[118] Dynamic DST Fusion (also known as DBF), was the highest-performing fusion algorithm for all of the classifier/detector combinations tested. Overall, this report provides initial evidence that joint human–CV systems have the potential to dramatically improve object detection. Work in related fields such as event detection and target tracking may also benefit from our methods.

4.2.2 Human-Centric Binary Classification Experiment

4.2.2.1 Overview

RSVP to humans has been explored in previous studies[72,157] and is fairly straightforward. Subjects are placed in front of a visual screen that presents a sequence of images, normally at a constant rate. The subjects are instructed to search for a specific target (or target type) and press a button when the target is

observed. Control over the pace of image switching allows the experimenter/operator to reach a desired balance between speed and accuracy.

The RSVP experiment used in the present study, summarized in Touryan et al.[130] was conducted with 18 human subjects to obtain EEG and button-press responses to images containing 0–4 "targets". The following are the number of targets per image: 1) 0 targets, 1347 images; 2) 1 target, 886 images; 3) 2 targets, 553 images; 4) 3 targets, 157 images, 5) 4 targets, 48 images, and 6) 5 targets, 9 images. An "office objects" data set of 3000 unique images was used in the presentations, with 5 different target types available: chairs, containers, doors, posters, and stairs. These images were collected and annotated by hand. In each experiment, subjects outfitted with EEG (BioSemi Active Two system, 256 channels) were notified of the designated target type and instructed to watch for targets. Subjects were also told to press a button if a target was observed. Note that the neural and BP data are used for classification, or categorizing full images as containing or not containing targets. The fraction of target to nontarget images ranged from 0.01 to 0.13 in each trial. Because of the small size of the image database, many of the same images (overwhelmingly nontarget images) were shown to the subjects multiple times; however, only the first instance of each image was used in the subsequent fusion experiment. The rate of image presentation was set at a constant 5 Hz (new image every 200 ms). Over a full set (6 trials), each subject was presented with approximately 17,000 images.

Trials were divided into 3 sets for each subject, one each for classifier training, fusion training (classifier testing), and fusion testing. Chronological order was maintained, and cross-validation was performed between different target classes in a specific temporal sequence to preserve set independence.

4.2.2.2 Neural Classification

Prior to neural classification, data was sampled and preprocessed according to Touryan et al.[130] The 256 channel dataset was down-selected to a subset of 64 that most closely matched the electrode locations in the standard BioSemi 32 EEG electrode arrangement. Offline, the EEG data were referenced to the average activity recorded at the mastoids, decimated to 256 Hz, and digitally band-pass filtered between 0.5 and 50.0 Hz. The 3 neural classification algorithms used in the present study follow:

- Hierarchical discriminant component analysis (HDCA): Ensemble method using temporally staggered logistic regression discriminators applied to 10 nonoverlapping windows in the first stage plus a separate logistic regression

discriminator applied over all first-stage discriminator outputs in the second stage.[59]

- XD+BLDA: XDAWN spatial filtering identifies a linear combination of the raw neural signals that maximizes the SSNR with regard to typical target and nontarget responses, then uses the weighted signal in a Bayesian linear discriminant classifier to calculate the final score.[61,78]
- Common spatial patterns (CSP) + BLDA: Spatial filtering method used to identify linear combinations of raw neural signals that maximize the variance between targets and nontargets.[158] Bayesian linear discriminant classifier calculates the final score.

These classifiers require a training step in which ground truth data is paired with actual responses, and the feature space dominated by target responses is separated from the space dominated by nontarget responses. In both training and testing, EEG data are broken up into epochs that correspond to per-image neural responses. These epochs are individually parsed by the NCs so that each image is assigned one score per classifier.

4.2.2.3 Button Press

A BP response to a target observation is a more concrete indication of an intentional human decision than EEG, which often contains high levels of noise. However, the timing of BPs relative to stimulus appearance is delayed and irregular, as shown in Sajda et al.,[157] typically occurring 2–5 images after stimulus onset (for RSVP at 5 Hz). Therefore, the likelihood of target presence in any preceding image is considered a function of the time between image onset and a BP. For each BP, images appearing in the preceding 1 s are assumed to have some probability of having caused the BP. As such, the BP "score" is calculated as the normalized difference between the reaction time and the median reaction time (empirically calculated from the training set).

4.2.3 Autonomous Object Detection

4.2.3.1 Object Detectors

Object detection is a quickly evolving area in the field of CV.[148] In our work, we apply 4 current algorithms based on graphical models:

- DPM: This method represents objects as a set of parts that can be deformed; using 2 different scales of HOG features, latent features, and a deformation cost.[127]

- SVM-based DSIFT: method based on matching densely sampled, pixelwise SIFT features between 2 images while preserving spatial discontinuities.[139]
- Exemplar SVM (ESVM): ESVM learns a separate classifier for each positive training image using a rigid HOG template and scores candidate detections based on "distance" to exemplars.[125]
- TAG: graph-based label propagation method using a small set of labeled images to derive likely labels based on image similarity metrics.[141]

4.2.3.2 Results and Discussion

Fusion of object detectors was conducted with and without augmentation by human-based classifiers, and performance comparisons were drawn between classifier/detector combinations, fusion methods, target objects, and human subjects. Similar to the analysis of object detection algorithms in related literature, we use average precision as the main indicator of detector performance.[148]

Table 10 documents average precision scores from individual CV-based detectors and individual human-generated classifiers. Among these detectors, DPM demonstrated the best performance (in every target category) followed by ESVM. While detection scores in current literature have surpassed these values (on different, but comparable datasets),[148] we reiterate that the fusion methods presented in this work are valid for any type of detector. The objective of this work is not to compete with state-of-the-art detection methods but to demonstrate that human and CV information can be combined to produce improved results, which will likely carry over to other detector combinations. In terms of human classifiers (here averaged over the first 5 subjects), BP yields the highest mAP, while XD+BLDA is the highest-performing NC. It must be stressed that average precision for a classification task cannot be directly compared with AP in a detection task, because detection requires localization while classification does not. Table 11 shows average precision scores using the 3 fusion algorithms on CV detections only. In terms of average precision, DBF is the best-performing fusion method in all target categories.

Table 10 Average precision, individual detectors, and individual classifiers

Detector	Chair	Container	Door	Poster	Stair	mAP
TAG	0.045	0.123	0.159	0.066	0.008	0.080
SVM	0.143	0.037	0.073	0.143	0.061	0.091
ESVM	0.125	0.318	0.150	0.236	0.122	0.190
DPM	0.188	0.396	0.194	0.342	0.143	0.253
Classifier	**Chair**	**Container**	**Door**	**Poster**	**Stair**	**mAP**
HDCA	0.229	0.135	0.152	0.165	0.195	0.175
XD+BLDA	0.346	0.224	0.234	0.226	0.289	0.264
CSP	0.183	0.106	0.134	0.145	0.185	0.151
BP	**0.476**	**0.366**	**0.332**	**0.346**	**0.426**	**0.389**

Table 11 Average precision, CV only

Fusion	Chair	Container	Door	Poster	Stair	mAP
Bayes	0.218	0.376	0.269	0.324	0.128	0.263
DST	0.198	0.318	0.163	0.273	0.124	0.222
DBF	**0.280**	**0.406**	**0.327**	**0.360**	**0.174**	**0.309**

Results of fusion output with the inclusion of human response are highlighted in Table 10. Here we focus on DBF because it was the best-performing fusion method. The first set of rows, incorporating CV detectors and all NCs (CV+NC) (but not BP [BP]) for 5 different human subjects, produces m scores (0.302–0.308) that are slightly lower than CV-only DBF (0.309). The second set of rows shows CV fused with BP (CV+BP), generating the second-highest mean scores (0.318–0.328) of any other set of combinations or fusion methods tested (CV+XD+BP was the highest [Table 12]). The third set of rows, which includes information from all detectors/classifiers (CV+NC+BP) actually produces lower average precision (0.301–0.310) than CV+BP. From these results, we can deduce that one or more of the NCs provides more contradictory than complementary information, leading to lower performance. Interestingly, target type did make a difference; for instance, DBF fusion of human information consistently improved "chair" detection, while it was consistently detrimental to "door" detection.

Table 12 Average precision, DBF fusion

			CV+NC			
Subject	**Chair**	**Container**	**Door**	**Poster**	**Stair**	**mAP**
1	0.287	0.391	0.313	0.374	0.159	0.305
2	0.313	0.386	0.332	0.361	0.144	0.307
3	0.335	0.368	0.281	0.375	0.162	0.304
4	0.265	0.411	0.312	0.380	0.170	0.308
5	0.282	0.439	0.312	0.351	0.127	0.302
			CV+BP			
Subject	**Chair**	**Container**	**Door**	**Poster**	**Stair**	**mAP**
1	0.322	0.399	0.331	0.377	0.167	0.319
2	0.320	0.396	0.316	0.367	0.210	0.322
3	0.317	0.397	0.319	0.392	0.195	0.324
4	0.273	0.432	0.320	0.385	0.229	0.328
5	0.331	0.422	0.322	0.358	0.159	0.318
			CV+NC+BP			
Subject	**Chair**	**Container**	**Door**	**Poster**	**Stair**	**mAP**
1	0.292	0.376	0.310	0.364	0.161	0.301
2	0.321	0.373	0.315	0.352	0.148	0.302
3	0.345	0.353	0.252	0.387	0.169	0.301
4	0.255	0.415	0.304	0.387	0.191	0.310
5	0.305	0.431	0.295	0.341	0.122	0.299

Table 13 compares the mAP for more combinations of detectors and classifiers and for each of the 3 fusion methods. It was noted that the XD+BLDA (XD) NC had much higher performance than the other 2 methods (HDCA and CSP). Therefore, fusion with CV+XD and CV+BP+XD was also investigated. In many cases CV+BP+XD produced higher results than CV+BP. Furthermore, averaged over the 5 subjects, CV+XD+BP yields the highest mean average precision. This provides strong evidence that combining BP and some types of neural classification can aid object detection.

Table 13 Mean average precision of different classifier combinations and different fusion methods

	***Bayesian Fusion*: CV+...**					
Subject	**NC+BP**	**BP**	**XD+BP**	**XD**	**NC**	**CV Only**
1	0.284	0.284	0.284	0.269	0.257	0.263
2	0.284	0.284	0.284	0.261	0.264	0.263
3	0.291	0.291	0.291	0.274	0.269	0.263
4	0.285	0.285	0.285	0.268	0.261	0.263
5	0.277	0.277	0.277	0.267	0.260	0.263
Mean	0.284	0.284	0.284	0.268	0.262	0.263
	***DST Fusion*: CV+...**					
Subject	**NC+BP**	**BP**	**XD+BP**	**XD**	**NC**	**CV Only**
1	0.206	0.221	0.211	0.202	0.201	0.222
2	0.204	0.226	0.210	0.196	0.195	0.222
3	0.205	0.237	0.208	0.202	0.203	0.222
4	0.214	0.239	0.224	0.222	0.212	0.222
5	0.218	0.221	0.227	0.221	0.222	0.222
Mean	0.209	0.229	0.216	0.209	0.207	0.222
	***Dynamic DST Fusion*: CV+...**					
Subject	**NC+BP**	**BP**	**XD+BP**	**XD**	**NC**	**CV Only**
1	0.301	0.319	**0.326**	0.325	0.305	0.309
2	0.302	**0.322**	0.314	0.313	0.307	0.309
3	0.301	0.324	**0.327**	0.324	0.304	0.309
4	0.310	0.328	**0.335**	0.324	0.308	0.309
5	0.299	**0.318**	**0.318**	0.315	0.302	0.309
Mean	0.303	0.322	**0.324**	0.320	0.305	0.309

To obtain a relative comparison of our classification performance in this detection task, a "perfect classifier" that correctly filters out all detection windows from nontarget images was simulated. When combined with CV detectors, an mPA of 0.385 was obtained for DBF, a difference of 0.052 from the highest fusion combination (Subject 4, CV+XD+BP). This puts the seemingly minor improvements obtained from the XD+BLDA and BP in perspective, suggesting the relative strength of the classifiers themselves and the limitations of classification in a detection paradigm. Coupling human EEG/BP response with gaze data (via eye tracking) would provide localization information to improve filtering of CV detection windows. This will be performed in future work.

Figure 26 shows the PR curves for different fusion algorithms, with 2 distinct levels of information (CV-only and CV+XD+BP) for 3 target types. As was indicated in Section 4.1.5.4, DBF outperforms DST and Bayesian fusion. In general, the addition of the XD+BLDA NCBPNC and BP response improves performance; however, the curves for CV-only and CV+XD+BP are similar in most

cases, suggesting that CV results dominate and accuracy of object detection. For practical applications, a real-time, hierarchical image-triage system could be set up to prioritize the most controversial/uncertain images to be inspected by a higher-level human operating at a much slower pace than RSVP.[88] Additionally, active interaction and active/proactive learning can be implemented in a real-time system.

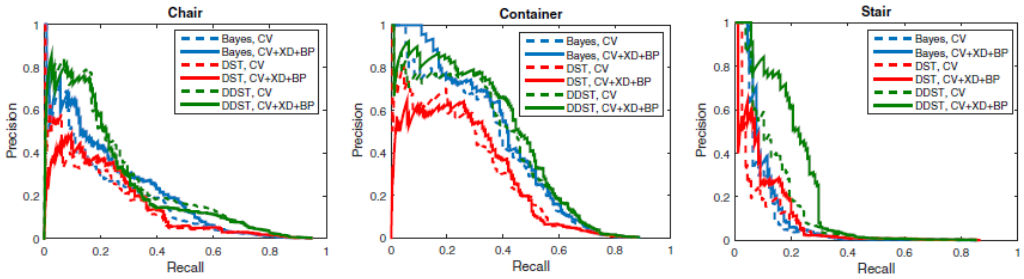

Fig. 26 Comparison of fusion methods and fusion combinations (CV-only vs. CV+XD+BP, Subject 4); multiple target types

4.2.3.3 Conclusions

A human–autonomy sensor fusion methodology was developed for the purpose of rapid object detection, demonstrating significant and consistent improvements in detection performance over CV algorithms alone across a range of different target objects. Of the various fusion algorithms and classifier combinations evaluated, Dynamic DST Fusion combining BP classification, XD+BLDA neural classification, and autonomous object detection yielded the greatest performance, indicating the potential of this new method. As late fusion methods are applicable to a wide range of information inputs, different human information and CV algorithms will be investigated and incorporated in future work.

4.3 Task Conversion

4.3.1 Introduction

Humans are unparalleled in their ability to recognize objects against complex or cluttered backgrounds. However, human perception is limited in throughput and may be substantially impacted by factors such as fatigue, boredom, and heavy cognitive workload. Furthermore, attempts to exploit human processing directly through the use of neurophysiological signals suffer a range of challenges that in most cases render them inferior to near-real-time graphics processing unit implementations of CV algorithms.[143]

In this report, we demonstrate that fusion of human detection with CV can enhance detection performance. We make use of the RSVP paradigm used by Touryan

et al.[130] In RSVP experiments, participants are instructed to press a button when target images are seen among images presented at a rate of 5 Hz. EEG signals are concurrently monitored for a signature that occurs after presentation of a target image, indicating a positive detection. The responses to RSVP denote the presence or absence of a target but do not identify the specific target type or localize it within an image. CV may be used to provide this additional information by applying specific target object models against location hypotheses in the image.

Using the unique conditions and capabilities of the 2 modalities, a family of algorithms was created to perform 4 related but distinct "tasks": determine 1) presence or absence of any target in an image, 2) presence or absence of a specific target type within an image, 3) presence or absence and location of a target within an image, and 4) presence or absence and location of a specific target type in an image. Figure 27 demonstrates queries given to the human and machine, responses, and fusion results for each of the 4 tasks. Because of the binary (presence or absence)-only nature of the RSVP paradigm, the human is only directly able to perform Task 1, while the computer can hypothetically perform all 4 tasks. (For this RSVP study, we treat incomplete results from Task 2 as complete results for Task 1. This has the effect of artificially deflating human performance figures but does not impact the validity of the fusion result. See discussion in Experiments and Conclusion, Section 4.3.5.) One way to mitigate this shortcoming of RSVP is to convert the information provided to a new form to encode the necessary information in an RSVP-compatible task. To that end, we introduce "task-conversion" strategies to allow for a meaningful human response in any of the 4 tasks and apply a combination of fusion approaches to exploit these jointly with CV detections.

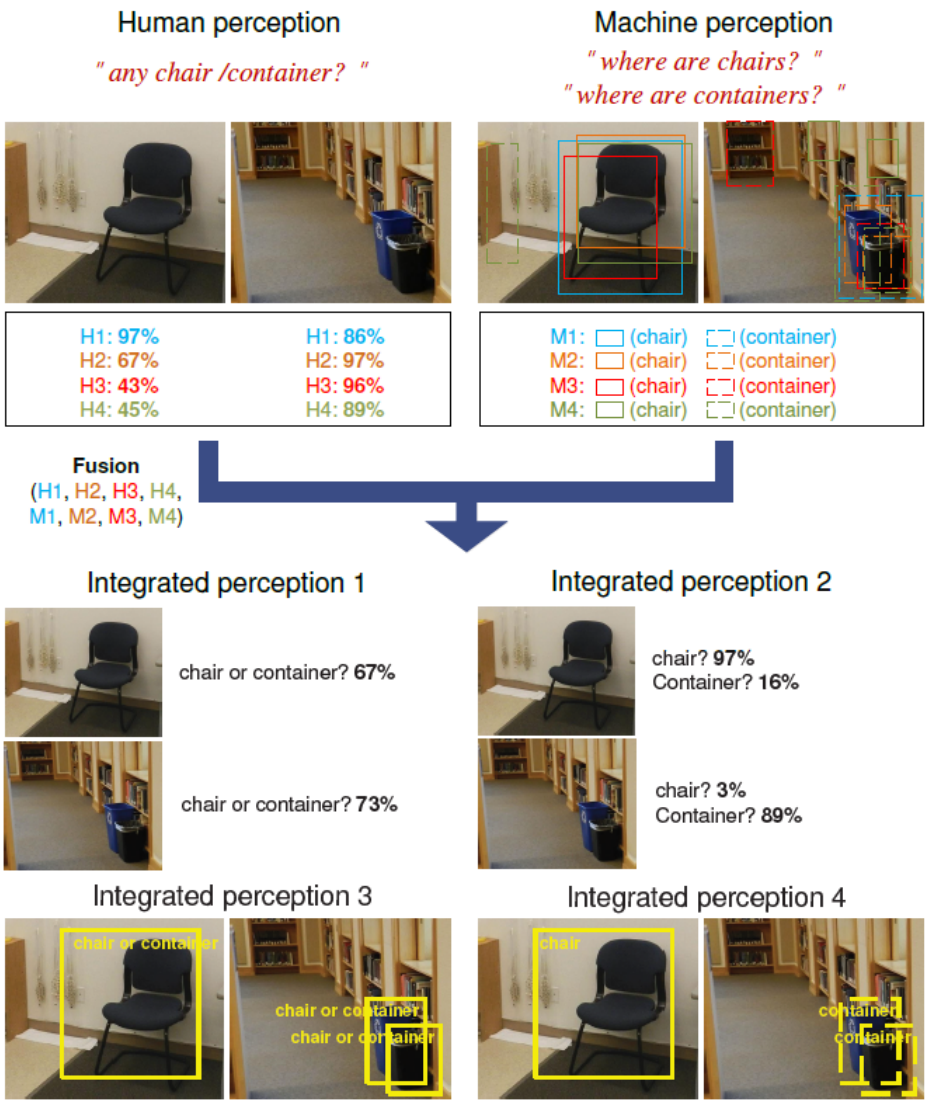

Fig. 27 Integration of human and machine perception in 4 different tasks: 4 approaches using human perception (H 1–4) and 4 approaches using machine perception (M 1–4). (top left) Query is given to the human subject. (top right) Machine perception results superimposed on image. (bottom) Fusion outputs of integrated human and machine perception.

4.3.2 Related Work

The RSVP paradigm traces its origins to the work of Potter and Levy,[159] who originally developed the approach to test the amount of information that human subjects could absorb at speeds that would not allow for deep periods of cognitive fixation. Early research was heavily biased toward text processing,[160] then there was a concurrent shift in latter decades toward practical applications in human-in-the-loop information processing. Mills and Weldon[69] proposed an RSVP-driven framework for dynamic text presentation that demonstrated mixed results versus other speed-reading approaches.

In the past 2 decades, there has been an accelerating body of research in the use of RSVP for human-in-the-loop image processing. The Defense Advanced Research Projects Agency's (DARPA's) Neurotechnology for Intelligence Analysts program[161] applied the technique, along with EEG processing, to the reduction in the search space of large amounts of previously unindexed imagery by human analysts.[59,62,162] Fei-Fei et al.[163] made the distinction between identification of targets in a scene, generally well correlated with the P300 EEG signal (onset at 300 ms after stimulus presentation) and the gist of images, which could be reliably detected in RSVP after a presentation of less than 100 ms. Evans and Treisman[164] found that when the contrast between foreground and background objects is strong, as with man-made objects with regular geometry against a natural background, with average time to detection in RSVP as little as 113 ms. Other groups[47,59,61,62,130,162] have demonstrated broad success in the use of RSVP to increase the throughput of human subjects analyzing imagery.

Recently, groups have attempted to integrate human-in-the-loop processing in a fully closed-loop system. Branson et al.[165] propose a "twenty questions" paradigm in which positive human response to one or more images in an RSVP set flags those images for downstream processing as a means of disambiguating between closely related classes of targets. Other attempts[166,167] at fused human–machine recognition systems generally interleave human interaction with an "active learning" phase, sequentially asking users to label examples in order to steer the underlying algorithm. Our previous work integrates human classifiers with CV-based detectors by a novel DBF approach in a heuristic way in object detection.[119] The task conversion we addressed in this report allows fusion in all 4 tasks.

4.3.3 Task Conversion and Fusion Strategies

4.3.3.1 Task-Conversion Techniques

The objective of this study was to effectively fuse responses from human and machine vision approaches. We identified the following 4 tasks that might be performed given human and machine responses to images, as illustrated in Fig. 28:

- Detect images containing any of the objects of interest
- Detect images containing a specific object of interest (for each object category)
- Detect location of any of the objects of interest
- Detect location of a specific object of interest (for each object category)

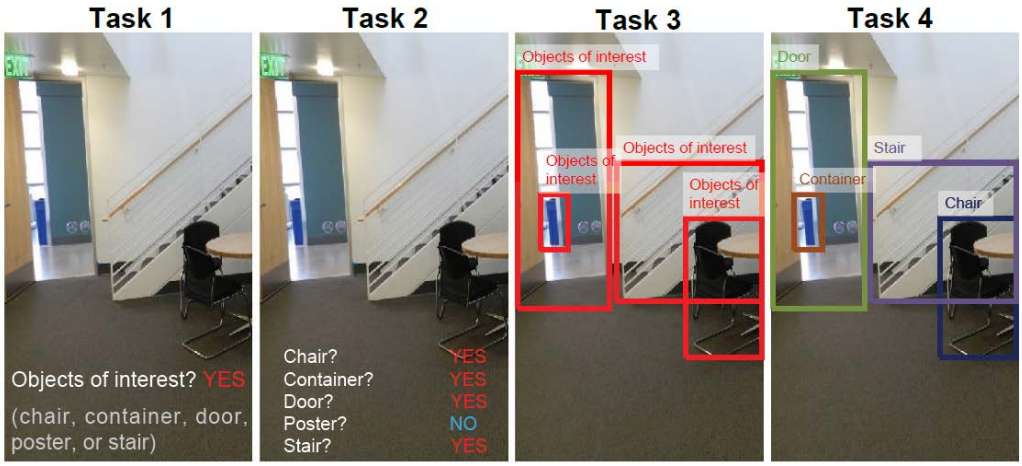

Fig. 28 Tasks 1–4 are categorizing images as 1) containing any of the objects of interest without localization, 2) classifying images for each object-of-interest class without localization, 3) categorizing images as containing any of the objects of interest with localization, and 4) classifying images from each object-of-interest class with localization

As previously stated, there is an inherent challenge in that human responses in this experimental design do not yield identity or location information for specific targets. The human is therefore only directly able to perform Task 1 while the computer can hypothetically perform all 4 tasks. In operation, object-detection algorithms generate candidate windows of various sizes and locations within the image and search for target features within each window. The output is a scored set of windows performed. When this procedure is done for each specific target class, it is analogous to Task 4.

Figure 29 demonstrates task conversion strategies for employing the output of CV-based object detectors in all 4 tasks. Initially, detectors search possible windows in the image for a particular object and assign each a detection score (i.e., Task 4). Each object of interest has an associated detector and the confidence scores of the detectors cannot be compared directly due to difference in scale (i.e., score range). We employ Platt scaling,[86] which calibrates the results of multiple detectors to allow them to be compared. Platt scaling, the process of rescaling and shifting the decision boundary of classifiers to create one unique boundary,[125] demonstrates that Platt scaling greatly improves the interdetector ordering while each decision boundary is no longer an optimal solution for learning each detector. Platt scaling learns parameters α and β, which are used in fitting a probability distribution of outputs of detectors to a shared validation set. (Here, the probability distribution is assumed to follow a sigmoid function with 2 parameters α and β.) The calibrated score f_s for the detection with detection score *sc* of the *i*th detector is as follows:

$$f_{s,(\alpha_i,\beta_i)}(sc) = \frac{1}{1+e^{a_i sc+\beta_i}}. \quad (17)$$

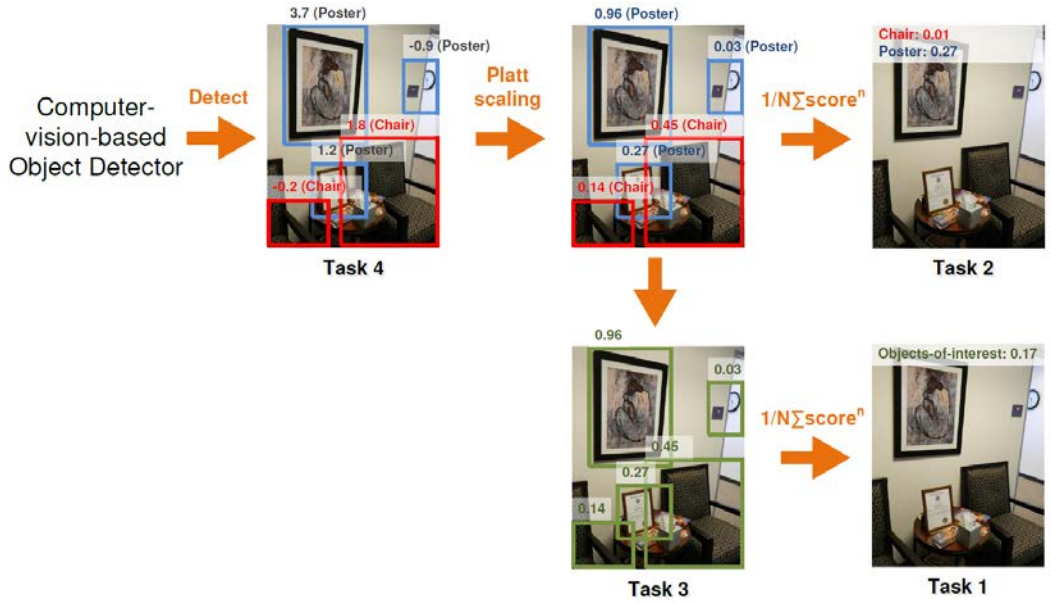

Fig. 29 Task conversion for CV-based object detectors

Details of implementation are described in 2 publications of the Institute of Electrical and Electronics Engineers.[118,119] Note that this generates a confidence score between 0 and 1 for each detector's candidate window(s), as in the top-middle image in Fig. 29, and preserves these scores when converting to a class-agnostic representation (Task 3), as in the bottom-middle image in Fig. 29.

We use the following aggregation formula to convert detection-level scores from Platt-scaled Task 4 (or Task 3) to a single image-level score in Task 2 (or Task 1):

$$score_I(H) = \frac{1}{N}\sum_{j=1}^{N} score_j(H)^k, \quad (18)$$

where $score_j(H)$ is the score of jth detection for a certain hypothesis H localized on image I, N is the number of detections in the image, and $k \geq 1$ is an empirical parameter. We choose a constant N for each image (15 in evaluation), resulting in equal reweighting across all images. If the original number of detections in an image is larger than N, only the N top-scoring detections are selected, and if the number is less than N, we consider the missing detections to be zero-scoring. We use the value of $k = 5$, which empirically proved to be effective in converting from a detection task to a classification task in Oquab et al.[142] Note that a higher k increases the contribution of high-scoring detections compared with lower-scoring detections. This aggregation formula is used for converting from scaled confidence for each object to Task 2 as well as from Task 3 to Task 1.

Figure 30 demonstrates task conversion strategies for the human response. It is impossible to directly infer object identity or location from the output of classifiers

because identification and localization are more difficult than pure classification. Similarly, estimating the location of objects of interest from the output of classifiers is also impossible. The performance of tasks requiring more-detailed inference should be worse, as in the sample PR curves shown in a right-most image of Fig. 30. (In a PR curve, greater area under the curve [AUC] denotes higher precision.) In a training step, we are able to identify classifiers and detectors that perform poorly compared with others. Here, we show that DBF effectively integrates human decisions with CV-based object detectors in all 4 tasks by treating the portion of performance caused by these factors as "uncertainty".

Fig. 30 Task conversion for human perception and precision and recall curve for all 4 tasks. Precision and recall are calculated for Subject 1's perception ability. For Task 2 and 4, the PR curve for the "chair" category is shown.

4.3.3.2 Integrating Outputs of Multiple Classifiers/Detectors

Figure 31 illustrates the strategy used to fuse multiple classifiers used in Tasks 1 and 2, in which each classifier produces only one score per image corresponding to the target hypothesis. To integrate the outputs of multiple classifiers, we fuse the image-level scores $\boldsymbol{s} = [s_1\ s_2\ \ldots\ s_K]$ where s_i is the output score of the ith detector. K is a number of classifiers.

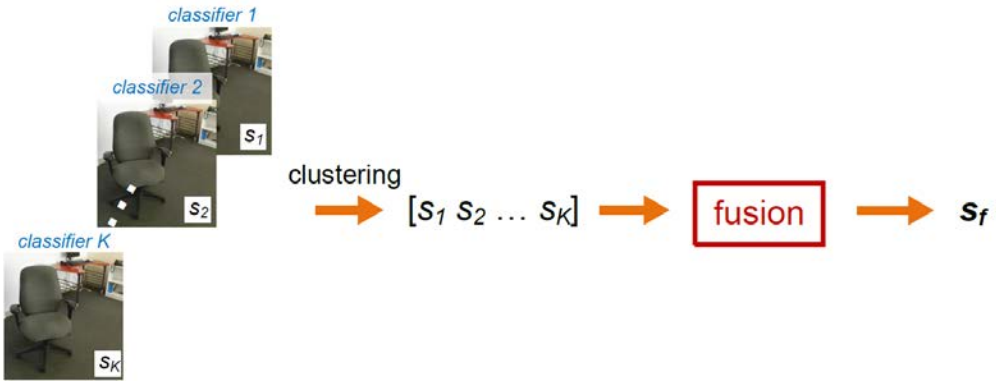

Fig. 31 Clustering process for Task 1 and 2

Figure 32 illustrates how detection-level scores (as produced in Tasks 3 and 4) from multiple detectors are fused. In Tasks 3 and 4, each detector outputs and scores

multiple candidate windows per image for the target hypothesis. We cluster detection windows that possibly contain the same target at the same general location and generate the score vector $\boldsymbol{s} = [s_{1\,j1}\ s_{2\,j2}\ \ldots\ s_{K\,jK}]$ for each cluster, where $s_{i\,ji}$ is the output score of j_ith detection window of an ith detector. Here, we consider 2 windows to be placed in the same location if the intersection over the union of their bounding boxes is over 0.5. Since clustering is performed for all individual windows of multiple detectors, multiple clusters corresponding to the same target hypothesis can exist. (For example, 3 clusters, each of which is detected in each image, contain the same chair in Fig. 32.) If a cluster contains more than one detection from the same detector, only the maximum score is inserted to the corresponding bin of the score vector. If a particular detector does not contain an overlapping detection window (of a particular target class) where others do, $-\infty$ is inserted to the corresponding bin, indicating no detection information is provided to the fusion from the detector. After calculating fusion score for all window clusters, we employ nonmaximum suppression to remove the clusters with lower fusion scores than other clusters in the same location.

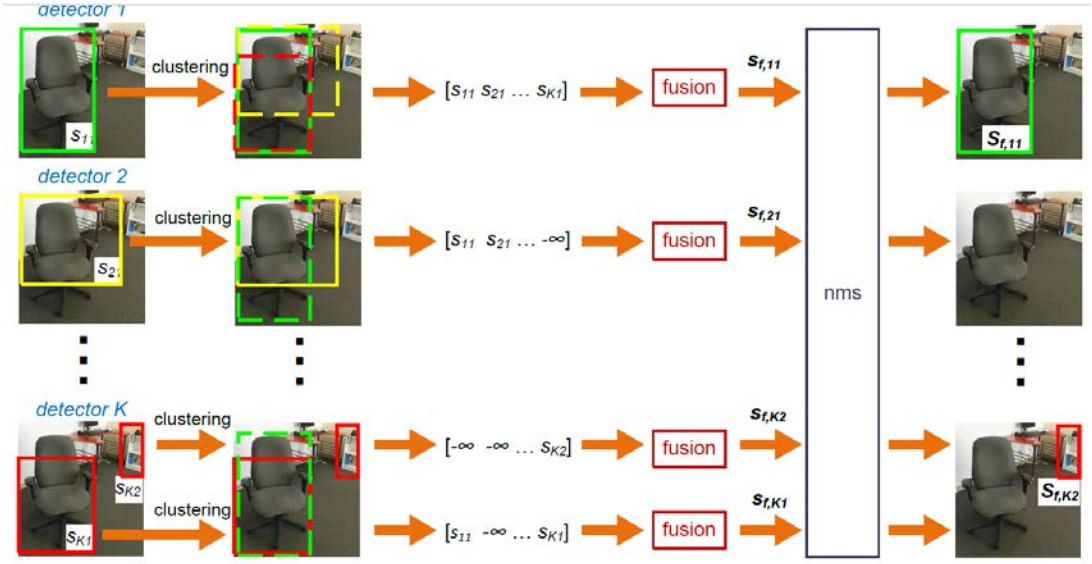

Fig. 32 Clustering process for Task 3 and 4

Fusion approaches: Naive Bayesian Fusion and Dynamic Belief Fusion

We employ 2 probabilistic approaches, naive Bayesian fusion[146] and DBF,[118] to create the final score vector $\boldsymbol{s}$.

Naive Bayesian Fusion assumes the classification and/or detection approaches to be fused are independent. The combination law, known as Bayes' rule, is given as

$$p(c|\boldsymbol{s}) \propto p(c)p(\boldsymbol{s}|c), \tag{19}$$

where c is a particular hypothesis, $\boldsymbol{s} = [s_1, s_2, ..., s_K]$ is the set of recognition scores from the constituent classifiers/detectors, and $p(c)$ and $p(\boldsymbol{s}|c)$ are the prior probability of hypothesis c and the likelihood of score $\boldsymbol{s}$ given the hypothesis, respectively. In this case, the possible hypotheses are presence and absence. These hypotheses can be applied at the image level (as obtained from the human RSVP response for any object category) or at the detection level (as obtained from machine perception for specific object categories). With the assumption of independence, a joint likelihood can be developed as the product of the likelihoods of K approaches:

$$p(\boldsymbol{s}|c) = \prod_{i=1}^{K} p(s_i|c). \quad (20)$$

The opposite hypothesis, $p(\neg c|\boldsymbol{s})$ can be calculated in the same manner. A model containing $p(c)$, $p(\neg c)$, $p(s_i|c)$, and $p(s_i|\neg c)$ was generated by aggregating scores in a validation set. During testing, prior and likelihood information were determined for each approach by referencing the model, and the final "fused" score was calculated as $p(c|\boldsymbol{s}) - p(\neg c|\boldsymbol{s})$.

Dynamic Belief Fusion[118] is an approach we previously developed to assign probability to hypotheses dynamically under the framework of DST.[128,129] DST is based on Shafer's belief theory.[128] Considering the 2 hypotheses, target (c) and nontarget ($\neg c$), it assigns probabilities that directly support those hypotheses and instantiates an intermediate state I, which represents evidence that could plausibly support either hypothesis. This intermediate state is given its own probability, quantifying the level of ambiguity that makes either hypothesis plausible. The belief function $bel(A)$ for a set A can be defined as

$$bel(A) = \sum_{B|B\in A} p(B). \quad (21)$$

If the probability is assigned to each of the 3 hypotheses, $bel(c) = p(c)$ and $bel(\neg c) = p(\neg c)$. [Note that $bel(I) = p(c) + p(\neg c) + p(I)$.] Once all classification and detection approaches assign probability to each hypothesis (including the intermediate state), Dempster's combination rule[129] can be applied to calculate a joint probability:

$$p_1 \oplus p_2(c|s_1, s_2) = {}^{1}/_{L} \sum_{X\cap Y=c, c\neq\emptyset} p_1(X|s_1)p_2(Y|s_2), \quad (22)$$

where $L = \sum_{X\cap Y\neq\emptyset} p_1(X|s_1)p_2(Y|s_2)$ and X and Y are subsets of 2^X. L is the sum total of probability mass whose common evidence is not the null set. Dempster's rule can be extended for multiple approaches using the associative and commutative properties of probabilities (i.e., $p_f = p_1 \oplus p_2 \oplus \ldots \oplus p_K$) with the following formula:

$$p_f(c|\boldsymbol{s}) = {}^{1}/_{L} \sum_{X_1 \cap X_2 \cap \ldots \cap X_K = c, c \neq \emptyset} \prod_{i=1}^{K} p_i(X_i|s_i), \quad (23)$$

where $L = \sum_{X_1 \cap X_2 \cap \ldots \cap X_K \neq \emptyset} \prod_{i=1}^{K} p_i(X_i|s_i)$.

Dynamic Basic Probability Assignment. In Fig. 21, the left plot shows a PR curve for an individual detector and a best-possible detector. The rates of values along the precision axis corresponding to recall r(s) are assigned as the basic probabilities to target, nontarget, and intermediate state, where s is a detection score. The right plot presents the basic probabilities with respect to a detection score, which converted from the PR curve.

For the *i*th classifier/detector, probabilities for a set of hypotheses {*c*, ¬*c*, *I*} are calculated using precision and recall information calculated in a validation set. For a given score *s*, the corresponding probabilities are given by

$$\begin{aligned} &p_i(\mathrm{c}|\mathrm{s}) = prec_i(s), \\ &p_i(\neg c|s) = 1 - prec_{bpd}(s) = rec_i(s)^n, \\ &p_i(I|s) = prec_{bpd}(s) - prec_i(s) = 1 - rec_i(s)^n - prec_i(s), \end{aligned} \quad (24)$$

where $prec_i$ and rec_i are precision and recall for the *i*th approach, respectively. $prec_{bpd}$ is the precision of a theoretical best-possible detector, which is assumed to have no ambiguity and is defined as 1 - $rec_{\mathrm{i}}(s)^{\mathrm{n}}$, where *n* is a parameter obtained by cross-validation and shared between all classifiers/detectors.[118]

After $prec_i$ and rec_i are computed in a validation set, testing is performed. Values corresponding to the test recognition score are used in the individual probability assignments for {c, ¬*c*, *I*}. Similar to naive Bayesian fusion, $p_f(c|\boldsymbol{s})$ - $p_f(\neg c|\boldsymbol{s})$ is used as the final "fused" score.

4.3.3.3 Approaches Analyzing Human Perception and Machine Perception

In the proposed fusion approach, we employ 3 NCs, BP, and 4 CV object detectors.

Approaches Analyzing Human Perception

Human visual perception can be estimated via biometric signals such as EEG or BP. For EEG, we employ 3 standard neural classification algorithms: HDCA,[59] XD+BLDA,[61,78] and CSP+BLDA.[158]

Each classifier is described in Marathe et al.[168] These classifiers require a training step in which actual responses are paired with ground-truth labels (target vs. nontarget) and the feature space dominated by target responses is separated from the space dominated by nontarget responses.

Compared with noisy EEG, BPs are a more concrete indication of a classification decision. However, when images are presented at a rapid rate, as is generally the case in RSVP, the timing of the BP relative to the stimulus presentation is usually delayed and irregular.[157] The likelihood that an image resulted in a BP is therefore some function of the time delay between presentation and BP. The BP classification score $s(t_p)$ at an image presentation time t_p is calculated by

$$s(t_p) = \left|t_p - (t_r - \Delta t_m)\right|, \tag{25}$$

where t_r and Δt_m are the BP response time and the median reaction time, respectively. $\Delta t_m$is empirically calculated from the training set. Figure 33 shows the BP classification-score computation.

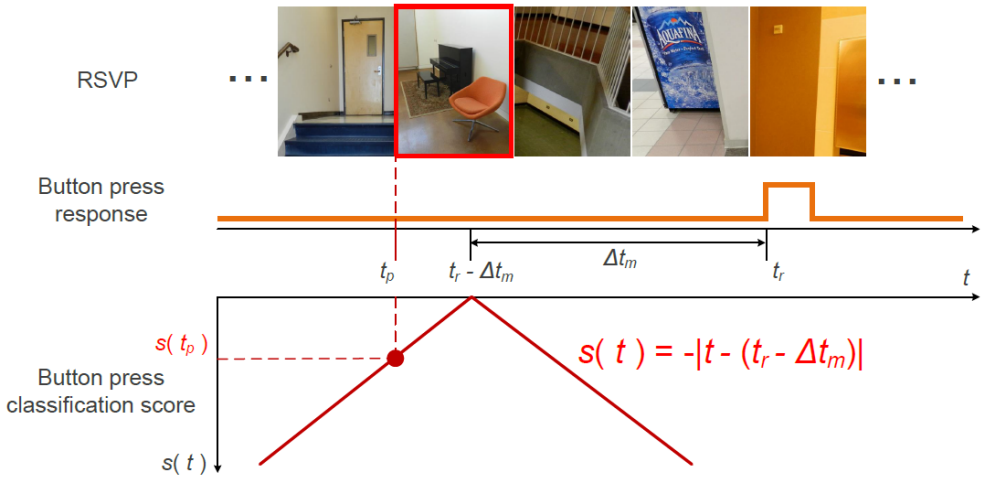

Fig. 33 BP classification-score computation: first and second rows demonstrate images presented by RSVP and BP response of participant when looking at a target, respectively; BP classification-score computation is shown in third row.

Machine-Learning Approaches for Machine Perception

The following 4 CV-based object detectors were selected for fusion with human decision:

- HOG+SVM)[126]: HOG features are employed to represent object appearance in terms of a distribution of gradients. An SVM is then trained to distinguish object from background.
- Exemplar SVM (ESVM)[125]: ESVM learns a SVM-based separate classifier for each positive training image (called as an exemplar) using a HOG feature, and scores candidate detections based on “distance” to exemplars.
- DPM[90]: Objects are represented as sets of parts that can be deformed using HOG features at 2 scales and latent features, with a deformation cost.

- FTCNN[142] is based on AlexNet[169] pretrained on a very large image data set, ImageNet.[170] The target-image data set used in this work contains much fewer images than ImageNet with quite different visual characteristics. To adapt the CNN structure of AlexNet to category distribution and characteristics of the target dataset, the final fully connected layer referred a classification layer is learned again over the target dataset.

Each of these techniques uses distinct principles for feature extraction and synthesis. We expect this will lead to the generation and fusion of complementary information.

4.3.4 Experiments

4.3.4.1 RSVP Data Set and Data Partition for Evaluation

The RSVP experiment used in this analysis presented images at 5 Hz (200 ms per image). The complete experiment consisted of 6 blocks of 10 min each (approximately 3000 images were selected for each block). Images in each block were randomly chosen but contained a specific ratio of target/nontarget images (this was a variable of interest in the preceding study). Six different ratios (0.01, 0.03, 0.05, 0.07, 0.09, and 0.11) were randomly assigned to the 6 blocks. Target images depicted at least 1 of 5 object categories: chair, container, door, poster, and stair. Because of the relatively small size of the dataset, images (mostly nontarget images) were repeated within blocks and between blocks.

Fifteen subjects participated in the RSVP experiments. The subjects were instructed to watch the sequence of images and to press a button when an object of interest was seen, for each target category (Task 2). Because of the incomplete performance of the task through the range of object categories, the results were ultimately treated as equivalent to Task 1. EEG data were collected in parallel using a BioSemi Active Two system with 256 channels (downselected to a subset of 64 channels that most closely matched electrode locations in the standard 10–10 arrangement), digitally sampled at 1024 Hz. Offline, the EEG data were decimated to 256 Hz and digitally band-pass filtered between 0.5 and 50 Hz. The neural and BP classifiers were trained and tested through a cross validation that preserved independence of these 2 sets in each cross validation.

In this experiment, partitioning of images for the RSVP task was random, with a fixed ratio of targets to distractors for each subject to enable the demonstration of the human–CV fusion concept. In subsequent work, we will further explore the effect of target to distractor ratios as well as individual choices of images by repeating multiple randomizations across subjects.

Unlike human perception, in which the response to an image can be altered by the subject's physiological state (fatigue, workload, etc.) or influenced by the preceding images, a (nonadaptive) CV-based algorithm will consistently produce the same outputs for the same image. As stated, CV algorithms cannot be trained, validated, or tested using repeated images due to the possibility of overfitting. Thus, the training, validation, and test sets were separate and nonoverlapping.

A specific data set partitioning procedure was performed to generate common validation and test sets, as illustrated in Fig. 34. The final 300 images from each RSVP test block were used as the fusion test set, as we hypothesize that the data at the end of each RSVP block elicit a steady-state subject performance level (to be validated in future work). The remaining RSVP images were collected to form a training set used for training the CV object detectors and a validation set used for computing the prior performance and likelihood model for each approach: essentially "fusion training". The partitioning between training and validation sets was done randomly at a 2:1 ratio. If repeat images from the RSVP task are not counted, the ratio among the train/validation/test sets was 2:1:2.

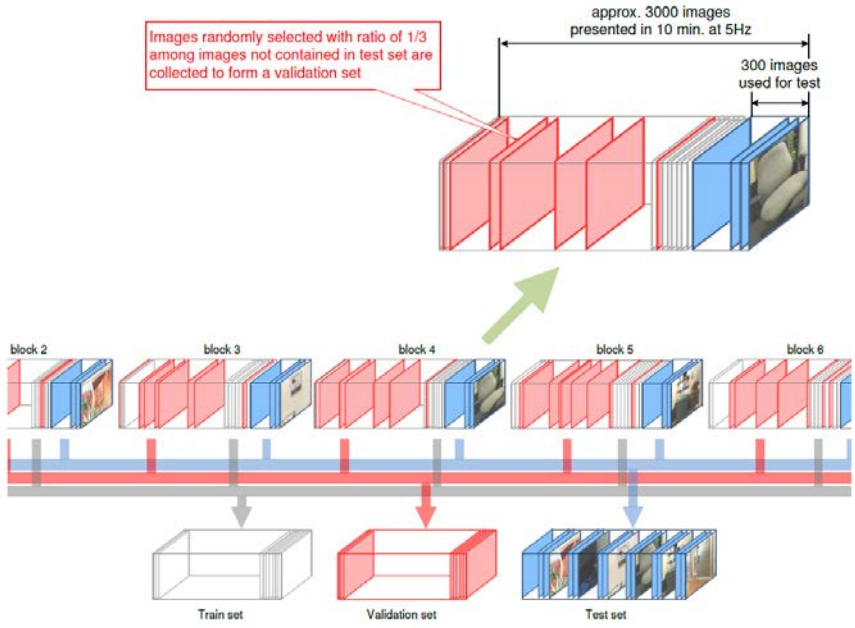

Fig. 34 Proposed partition of the image set used in the RSVP task: image set consists of 6 blocks; for each block, the last 300 images are used for testing. Images not contained in test set are randomly split into training and validation sets at a 2:1 ratio.

4.3.4.2 Performance Comparisons

Performance was evaluated using AP for Tasks 1 and 3, and using mAP across object categories for Tasks 2 and 4. AP is a standard metric in the CV field, obtained by averaging precision values across the range of recall; in that sense, it is semantically similar to the AUC metric. AP/mAP results from naive Bayesian fusion and DBF were calculated and averaged over all subjects. The fusion performance of human-only, machine-only, and human+machine (i.e., "all") perception are shown in Fig. 35.

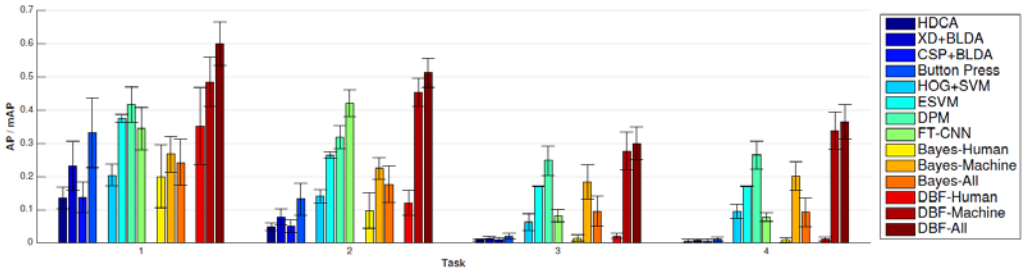

Fig. 35 Performance comparison of individual approaches and fusion approaches (Bayesian fusion and DBF). For each fusion approach, 3 bars indicate results of integrating of human-perception approaches only, CV-based approaches only, and all perception approaches. Fusion is performed in 4 tasks and results are shown in order; error bars denote standard deviation across subjects.

For all tasks, we employed one-way ANOVA tests to assess the effect of fusion method on AP/mAP. Each one-way ANOVA considered the choice of individual approaches (e.g., DPM) or fusion approaches (e.g., DBF-machine) as a main effect (14 approaches total). The results across 15 subjects are as follows: [$F_{(13,182)}$ = 73.87, $p < 0.001$], [$F_{(13,182)}$ = 375.27, $p < 0.001$], [$F_{(13,182)}$ = 237.39, $p < 0.001$], and [$F_{(13,182)}$ = 378.41, $p < 0.001$] for Tasks 1, 2, 3, and 4, respectively. These results imply that, at the very least, there were statistically significant differences between the best and worst approaches in each task. As a follow-on test, we compared the statistical difference between any pair of 14 approaches by using a multiple comparisons test. Although the statistical difference between DBF-machine and the best individual approach (DPM or FTCNN, depending on the task) is not significant in Tasks 1–3, by fusing human and machine perception (DBF-all), performance does yield statistically greater performance than all other approaches. Note that this occurs even in Tasks 3 and 4, where human-only performance with DBF is very poor. These results support our hypothesis that human and machine perception yield complementary information that can be leveraged for improved performance. Bayesian fusion is statistically lower-performing than the best individual approach in all 4 tasks, is negatively influenced by poor-performing approaches (i.e., human perception), and cannot adequately resolve conflicting information. Furthermore, the performance of human+machine Bayesian fusion consistently drops below

machine-only fusion; the opposite is true for DBF. This supports the key concept behind DBF, that modeling an intermediate state to represent "uncertainty" in detector outputs can be beneficial. The fact that these results are consistent across all tasks suggests that the structure of tasks themselves do not affect the outcome.

The results of each task cannot be directly compared with one another. For instance, Task 2 is more complex than Task 1; thus, it is understandable that AP/mAP will be lower. However, it is interesting to see that DBF-machine and DBF-all in Task 4 (target-specific) outperform Task 3 (target-agnostic). This may be because the task-conversion algorithm from Task 4 to 3 did not adequately convey the relative certainty between target types (e.g., a score of 0.27 for a poster target may not actually represent the same confidence level as a score of 0.27 for a chair target). The 0.02–0.03 mAP performance loss during task-conversion of certain individual approaches (ESVM and DPM) caused a significant performance loss in fusion (0.07 for DBF-machine and 0.08 for DBF-all). Pairwise t-tests indicated a statistically significant difference ($p < 0.001$) for each pair of tasks, including Tasks 3 and 4.

To evaluate the effectiveness of fusion between subjects, we compare the 6 fusion approaches against the best human-based classifier and machine-based detector for each of the 15 subjects for all tasks in Fig. 37. For all except the fifth subject in all tasks, DBF-all demonstrated the greatest performance. For Task 1, DBF-human outperforms the best human-based classifier 12 of the 15 subjects and DBF-machine outperforms the best machine approach for all subjects. As suggested by the previous analysis, the combination of human and machine information using Bayesian fusion underperforms the best individual human and machine approaches for all subjects. The comparisons in Tasks 2–4 shows similar tendency as in Fig. 36.

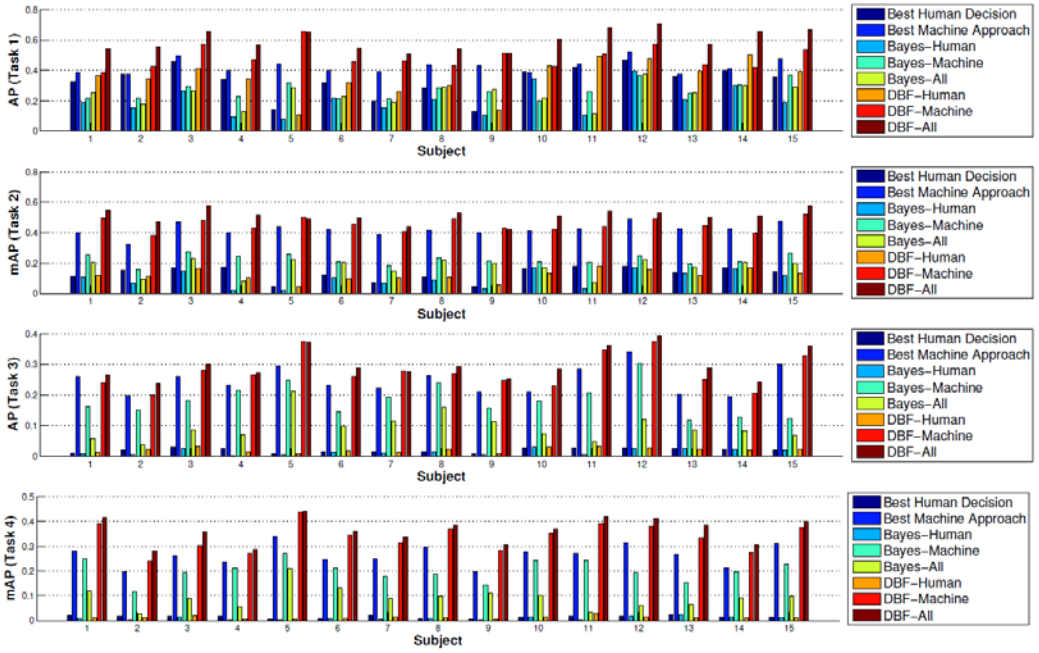

Fig. 36 Performance of best human and machine approaches as well as fusion approaches per subject

One possible confound in using mAP as the metric for comparison is that the measure of precision does not incorporate the number of "misses" for a given classifier. For an object detection task (Tasks 3 and 4), this type of metric is appropriate; however, for a classification task (Tasks 1 and 2) the misses are often just as important as the accuracy of hits. Thus, to verify that the performance differences we saw using mAP for the classification tasks were not simply a product of the chosen metric, Tasks 1 and 2 were also evaluated using AUC, a conventional metric for classification tasks. A comparison of AUC for all individual and fused approaches is shown in Fig. 37. DBF-all outperforms all individual classifier/detector approaches as well as all fusion approaches, as was observed using the AP/mAP metric. DBF-machine and DBF-all fusion also outperform all versions of Bayesian fusion. A one-way ANOVA test was performed on these results with fusion approach as the main effect, obtaining [$F_{(13,182)} = 35.54$, $p < 0.001$] and [$F_{(13,182)} = 66.65$, $p < 0.001$] for Tasks 1 and 2, respectively. Subsequent multiple-comparisons tests demonstrated that DBF-all yields statistically superior mean performance over all approaches except DBF-machine and FT-CNN (the best CV approach). Note that DBF-all outperforms all other approaches in 13 of 15 and 11 of 15 subjects for Tasks 1 and 2, respectively, indicating a strong trend toward the superiority of DBF-all. We conclude that generating DBF likelihood models provides superior fusion performance based on a combination of the mAP and AUC performance metrics.

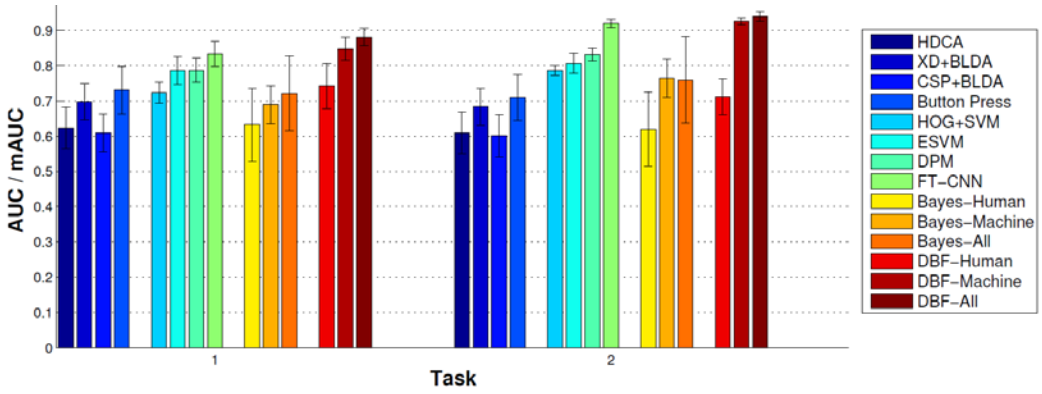

Fig. 37 Performance comparison with AUC: error bars denote the standard deviation across subjects

4.3.5 Conclusion

We have shown that combining RSVP with CV algorithms in a fusion framework can substantially improve target recognition performance using the precise localization and identification capabilities of CV in combination with the precise detection capabilities of a human operator. We have specifically leveraged perception feedback detected from an EEG signal in combination with BP response, applying positive detections to 4 state-of-the-art CV approaches. Additionally, we have implemented a unique task conversion protocol that overcomes some of the limits of human detection via RSVP, namely the lack of localization and identification feedback.

The treatment of incomplete results for Task 2 as being equivalent to Task 1 has the effect of artificially deflating performance numbers for the human subject response. However, fusion via DBF is still able to leverage unique information from both sources in order to increase overall detection performance. These results suggest that human performance in high cognitive workload situations may be enhanced by integration with fusion approaches, including a novel DBF approach. Future work will further explore the effect of randomization of image sequences across subjects and the effects of variable human-subject performance over time.

5. Agent Adaptation

In adaptive systems, dynamics in both the environment and in human performance can lead to a breakdown in performance for autonomous systems. Here we focus on adaptations to dynamics in the environment. In these situations, most algorithms fail, catastrophically, the first time they are exposed to more-complicated real-world scenarios. These include the effects of variable lighting and nonstatic backgrounds, such as waving trees, and state estimation errors due to mobile agents

present in the field of view, such a Soldiers completing a mission. As we anticipate this problem occurring, we have taken steps to mitigate the problems will know are coming. We describe a study that address the issues of dynamic lighting and dynamic scenes.[171]

5.1 Dynamic Lighting

This project was motivated by time our team spent embedded in an infantry unit in a forest in Columbus, Georgia. The Soldiers requested an unmanned aerial vehicle (UAV) that could be pulled out of a pack, turned on, and then fly through the trees, find a clearing, exit the canopy, surveil the area while transmitting video data back to the ground station, reenter the canopy, and return to the operator. We have chosen to address the forest environment such a UAV would operate in. Forests are highly complex environments. In addition to natural variations in illumination due to passing clouds, the trees create irregularly placed, occasionally mobile beams of light. Furthermore, we are constrained to a theoretical small-scale, easily man-packable UAV with an extremely limited payload that must move at operational tempo, currently defined by DARPA as 20 m/s.[172] We anticipate this platform will have the bare-minimum number of sensors, possibly only a single monocular, grayscale camera, and we have designed this system to be robust to such a possibility.

To enable stable flight on pocket-sized, highly dynamic unmanned aerial systems (UASs), the control loop requires extremely high update rates, but such platforms have only a minimal payload to handle the computational burdens. As system size decreases, high-fidelity sensors, such as LiDAR, become too large to carry, requiring a shift to smaller sensors for state estimation. Optical systems, which easily scale down, typically use either stereo vision or optical flow to generate state estimates. Both methods rely on the objects in the scene maintaining a static representation for correlation, which causes them to be susceptible to dynamic lighting-induced errors. Even large systems that can carry LiDAR and substantial computation often couple high-fidelity systems with optical systems to increase their effective range.[173] While there are many optical navigation techniques available, we have chosen to limit the scope of this problem to optic flow due to its low computational burden and proven real-time usability.

Traditionally, optic flow, including variants of Lukas–Kanade,[174] elementary motion detectors,[175] and Horn and Schunck,[176] has been tied to a static representation of the world. All of these algorithms assume that color representations remain consistent from one frame to the next, allowing perceived motion to be tracked by observing pixelwise changes between the 2 images.[176] This

assumption becomes a problem when lighting conditions change unpredictably, a fairly common occurrence outdoors, as it becomes nearly impossible to track the motion of individual pixels that have now lost their primary signature. As lighting regularly ranges from 80 lux indoors to 1000 lux outdoors on an overcast day to more than 130,000 lux in direct sunlight,[177] this is a substantial problem.

5.1.1 Related Work

Before we address similar research, we would like to mention auto-exposure and similar in-system camera settings that deal with varying lighting conditions. While such adjustments, including histogram shifts and other sorts of calibrations, may produce an image that is similar and comprehensible to the human eye, the actual pixel values (as represented from 0–255 in a grayscale image) often differ beyond what is recognizable for an algorithm making a correlation from frame to frame.

While there have been numerous explorations into illumination invariance, there remains no standard method to improve the performance of optic-flow systems in dynamically lit, complex, novel environments. Illumination mitigation techniques run the gamut from computationally expensive postprocessing using a variety of filters and masks on the imagery to preprocessing to alter the imagery before it goes into the optic-flow algorithms.[178–192] Some, such as Dederscheck et al.,[178] work in highly constrained environments, such as highways, and assume that the light will shift evenly over the entire image.[178,179]

As our system is intended for a forested environment, such constraints were felt to be unrealistic. For similar reasons, we move past methods that perform object identification, and the comparison of the colors found to those of a template were used to estimate the lighting changes and correct the image as a whole.[180–182] We do not wish to correct the image but rather simply use it to compare against another image. Understanding what it represents is not within the current scope of our investigation.

Much closer to our problem space are methods that look to buttress optic flow estimation through the use of additional sensing modalities, such as LiDAR,[183] inertial measurement unit (IMU),[184] or temperature scans.[185] While we acknowledge the utility of these methods, they do not address our chosen problem of unaided optic flow. A small-scale UAV is highly constrained by its payload capacity and processor and such methods are often infeasible. In Zimmer et al.,[186] the authors investigate an implementation of optic flow in HSV (hue saturation value) color space, where they found the hue channel is invariant under a variety of illumination changes and does not show marked reaction to shadows or specularities. We feel this is a very interesting avenue of investigation, but it was

not utilized, as our target platform uses grayscale cameras because we are attempting to reduce the computational burden, and tripling the number of pixels involved would not further that goal. The authors would also like to differentiate this technique from gradient-based optic flow.[187,188] Our method preprocesses the input to any optic flow algorithm but is not one itself.

The techniques most similar to our approach look at preprocessing the input prior to putting it through an optic flow algorithm. Christmas[189] shows the use of a spatial filter to reduce the effects of temporal aliasing on image pairs with a large discontinuity. This work found that in an experiment with constant illumination and even motion pattern it was possible to reduce temporal effects using a low-pass filter. However, due to the sizes of the filters investigated, they predicted difficulty in real-time applications.

In Lempitsky et al.[190] the authors removed the effects of shadows by subtracting from the first image, the result of that image was convolved with a Gaussian kernel. Beyond the fact that this work continues in RGB color space and is finally computed with bicubic interpolation, which is far more computationally complex an approach than we will be able to use for our application, is the fact that this technique ignores higher-order variations. Sellent et al.[191] come at the problem from a completely different angle, choosing to deal with natural variation in lighting by extending exposure time and using the camera itself to prefilter sharp variations out of the imagery. While this approach is not practical during flight, it is a clever approach for static platforms. In Sharmin and Brad,[192] images are filtered prior to use in an optic flow algorithm, and like the work presented in Sellent et al.,[191] they have used a Gaussian smoothing filter, although that paper presents a filter tuned specifically for a Lucas–Kanade implementation at each pyramidal level, and it too suffers from computational complexity limiting its real-time applicability.

Our investigations have found that, unlike standard electro-optical (EO) imagery, the double derivatives of those same images remain fairly stable in dynamic lighting conditions. Image derivatives provide the rate of the change of local intensity; in other words, highlighting the edges and corners. Whether the scene is brightly or dimly lit will not change the textures of the objects in the scene, and barring extremes, such as near-darkness or imager saturation, things like tree bark look more or less the same no matter how they are lit. However, with image derivatives, there is an overall reduction in the amount of information available. Fine details disappear, color is removed, and boundaries that lack a sharp textural difference can merge into one. While our approach has been discounted in prior work due to this loss of information,[183] we find that there is sufficient complexity in realistic outdoor environments to maintain a high enough information content to

allow navigation and control using only the double-derivative image. Even in relatively sparse environments such as images with large tracks of sky, where the estimation uncertainty grows, there is a sufficient quantity of information that our method does not experience the catastrophic failures associated with other optical methods.

5.1.2 Methodology

The method presented here is intentionally very simple. Small-scale aerial platforms do not have the capacity for complex, real-time calculations while flying at Army-desired operational tempos.[171] There are 3 steps to our method: 1) solve for the double derivative, 2) apply optic-flow algorithm of choice and, 3) isolate the true flow using a mask, as shown in Fig. 38.

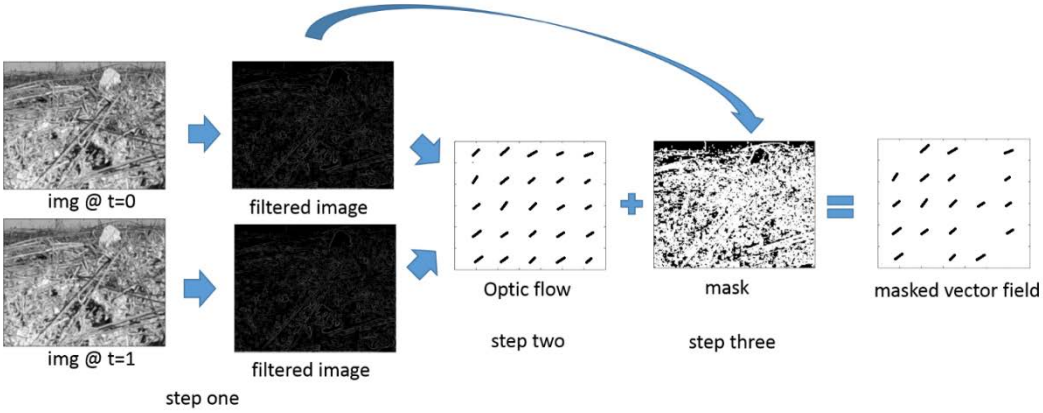

Fig. 38 Flow chart of proposed method, showing the 3 steps employed in our method

Standard derivative calculations were used, incorporating both the x and y components. While several different filter operators were tried during the course of the project, we chose [1,- 1] (or its transposed pair, for y). Numerous other standard filters (Gaussian, Laplacian, etc.) were tried and produced no significant increase in performance. Due to the continued focus on reducing the total number of computations required, a smaller filter operator was preferred.

$$\nabla f = \left(\frac{\partial f}{\partial x}^2 + \frac{\partial f}{\partial y}^2 \right)^{-2}, \tag{26}$$

where $f =$ is the value of the pixel at the (x,y) position, and the second spatial derivative of f is the gradient of the image (∇f), as shown in Eq. 26.

Those areas of the image without any texture will result in no detected motion from optic flow algorithms. To avoid biasing the state estimation, and to still allow for the possibility that there may not have been any ego-motion, we created a mask to

put over the resultant optic-flow vector field. This mask is based on the information-poor areas of the filtered image. Through experimentation, we found that masking all areas where the double derivative image pixel value was below 10 (on a scale of 0–255) removed the majority of the spurious data.

Regarding complexity, let $n = w \, x \, h$ be the number of pixels in each image frame. We assume the use of optic flow algorithms to which the input consists of sequential pairs of frames in a time sequence. In the case of Lucas–Kanade, the time complexity of each iteration of optical flow is dominated by the computation of the Hessian, which is linear in n but quadratic in the number of warp parameters m ($\approx O(nm^2)$).[193] In the case of iterative global methods such as Horn–Schunck, analysis of the computational cost has additional dependencies including whether the algorithm is allowed to converge. At its core, the Horn–Schunck algorithm is the Jacobi iterative method applied to the interior of the image[194] and requires computation of first-order partial derivatives with a complexity at least linear in n.[195] A highly optimized implementation may precompute certain pixelwise quantities prior to entering the iterative phase. Even so, it has been demonstrated practically that Horn–Schunck requires on average substantially more computations per pixel than Lucas–Kanade.[196] Our technique is a preprocessing step requiring a single convolution over each image to be taken as input to the chosen optic flow algorithm. The results of the derivative computation may be stored for later use at a cost no greater than that of the input image. Let k be the size of the convolution kernel ($k = 2$ in the method suggested previously). We assert that the complexity of our method is linear in n, and the subsequent masking operation requires constant time per pixel, resulting in an overall complexity $O(nk + cn)$. We posit that this complexity is generally dwarfed by that of the downstream optical flow algorithm.

The implementation of Lucas–Kanade[197] employed for these experiments used 3 pyramids and 3 iterations, while that of Horn–Schunck[198] used an alpha of one and maximum iteration of 100. In both cases the algorithms were received pretuned and were not altered.

The authors are aware that this method is not robust to the “features” created by strong shadows. However, from our experience embedded with infantry units in realistic conditions and the point-of-view video data gathered during those experiments, strong shadows remain relatively static, and we are more concerned with dynamic changes such as the passage of clouds overhead, which will affect all shadows and lighting conditions equally. However, we have begun investigation in texture patterns which may mitigate the shadow problem.

5.1.3 Data

The images used in the following experiments were collected in a field at Fort Benning, Georgia, during a live exercise[199] using a GoPro Hero3 and in an office using a Logitech C920 Webcam. Stage lighting was used to create controllable and repeatable dynamic lighting conditions for the indoor space. The outdoor data collection occurred on a cloudless and extremely sunny day, and its lighting is considered to be static.

For ease of comparison, both datasets were collected with a static camera. To create the companion "dynamic motion" sets, the "static motion" sets were subsampled and only a portion of each successive frame was used. The location of this region within the base image was altered by a known number of pixels, which became the ground truth for the optic flow estimations. To ensure that the data were comparable, the static sets used an identically sized subsampled area from the original images, although in this case the location of the window did not alter between frames. Each dataset comprises 100 sequential images, selections of which are shown in the videos in the following subsection.

There have been several data sets created to test visual-state estimation under a variety of circumstances, most notably the KITTI[200] and Sintel[201] data sets. Both include a variety of lighting conditions; however, we did not find known illuminance values associated with the imagery. As this is preliminary work in which we hoped to explicitly measure the relationship between changes in illumination and the amount of noise added to an optic flow estimate, we chose to create our own data set. In future iterations of this work we plan to use established datasets to allow for rigorous comparison between methods.

5.1.4 Experimental Scenarios

In the following scenarios, the term "algorithms" refers to both Horn–Schunck and Lucas–Kanade with both standard and prefiltered inputs.

Scenario 1: Static Camera, Static Lighting

To ensure that any error was due to the input, rather than the algorithms, we first tested the performance of the optic-flow methods using the statically lit data set and a static camera location. With no motion and no sensory noise, an ideal output would show a cluster of dots around the (0,0) point.

Scenario 2: Static Camera, Dynamic Lighting

In this case, we wished to measure how the algorithms responded to changes in illumination only. With a static camera and static scene, any resulting "motion" must be due to the illumination shift.

Scenario 3: Dynamic Camera, Static Lighting

To determine how accurately each algorithm responded to actual motion, we tested them using a dynamic camera position and static lighting condition. In this scenario, any deviation from ground truth must be due to errors in state estimation of the algorithms themselves.

Scenario 4: Dynamic Camera, Dynamic Lighting

The final scenario investigates a mobile camera in dynamic lighting conditions. We believe this is the most similar to true field-operation conditions and shows how accuracy of the algorithms' state estimation. In this case, deviation from ground truth is due to a combination of motion and lighting conditions.

The results of the first 2 scenarios may be found in Fig. 39 and the last 2 in Fig. 40.

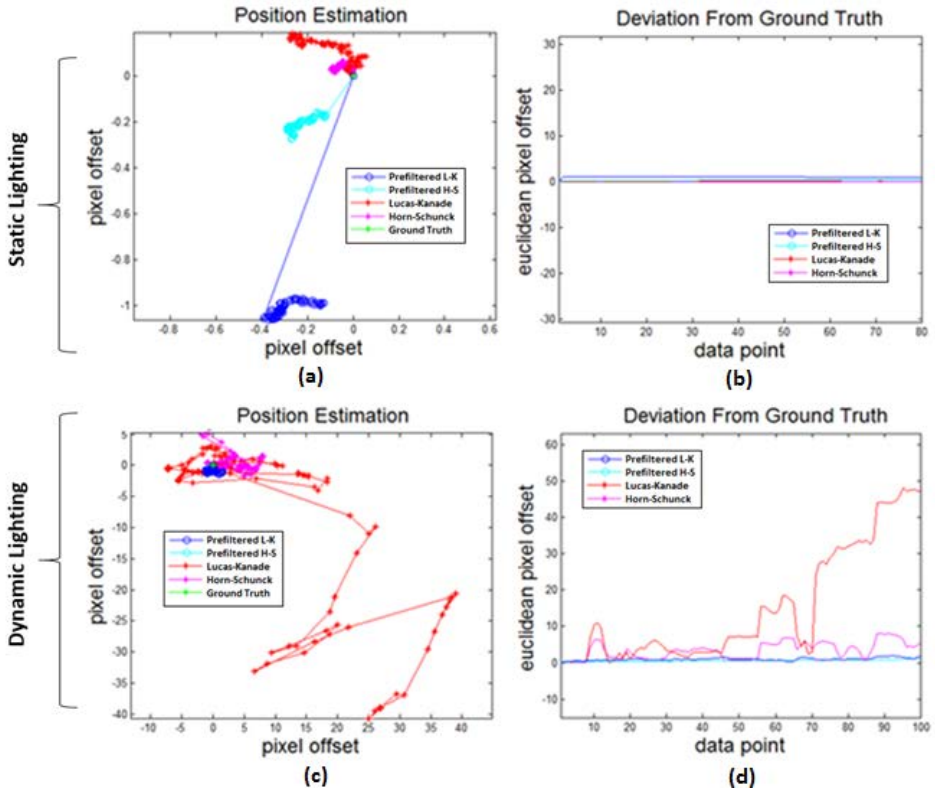

Fig. 39 Measured responses for a static camera position and static picture with dynamic changes in lighting, where a) and b) show the position estimation and deviation from ground truth for Scenario 1 and c) and d) show the results of Scenario 2

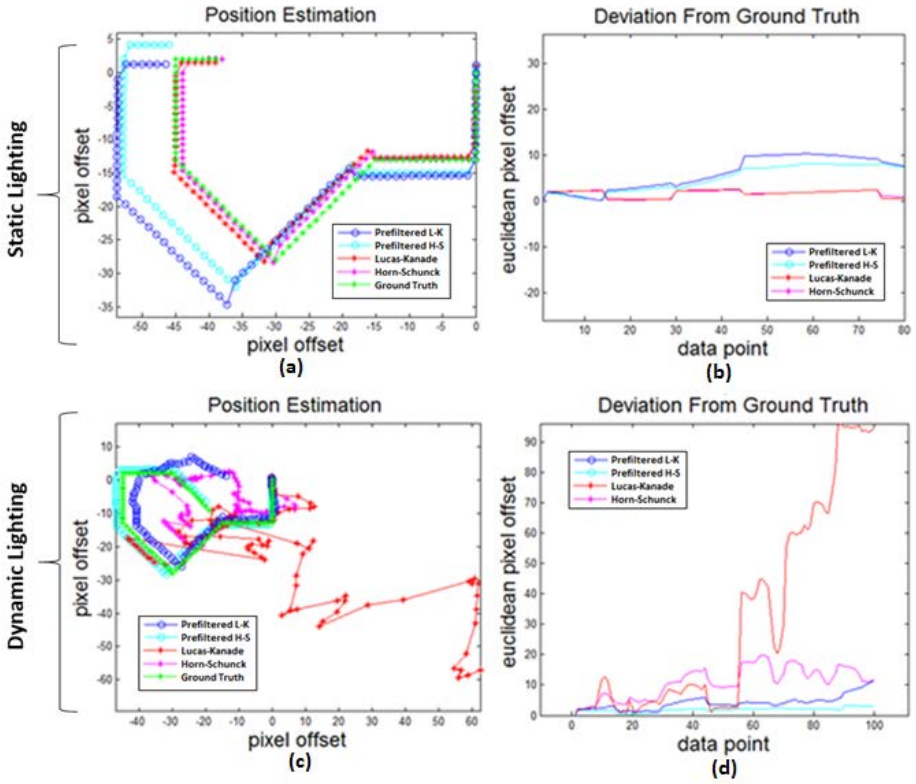

Fig. 40 Measured responses for a dynamic camera position and static picture with dynamic changes in lighting, where a) and b) show the position estimation and deviation from ground truth for Scenario 3 and c) and d) show the results of Scenario 4

5.1.5 Results

The following graphs represent the results of processing the aforementioned lighting conditions with and without the prefiltering step. In each figure, the same camera path is shown in both static and dynamic lighting conditions. By showing how the algorithms respond to their expected input (statically lit images) as well as how they respond to the dynamic lighting, the reader may see which errors are due to the implementations of the algorithms and which are due to the input data format. Furthermore, we include the same implementation with and without the prefiltering step, so that the reader may directly see the difference it makes. The left column of each figure shows the actual plotted position of each state estimate as well as the true position, while the right column show the Euclidean disparity between the estimate point and ground truth at each position in the run. Deviation is measured in pixels.

Figure 39 shows the results of static (a,b) and dynamic (c,d) lighting on optic flow using data collection with a stationary camera to demonstrate the error added to state estimation through variation in lighting alone. The camera is static; any perceived “motion” must be due to estimation error. As may be seen in the top row,

all algorithms perform as expected, showing subpixel motion when presented with no motion and static lighting. However, in the bottom row, one may see the results of varied lighting conditions and that the filtered implementations are far more robust it effects.

Figure 40 shows the results of optic flow in both static (a,b) and dynamic (c,d) lighting conditions as captured by a dynamic camera. In this case, motion may be due to either actual change or lighting, although as may be seen by comparing the top (static lighting) and bottom (dynamic lighting) rows, all algorithms perform well in static lighting, while the prefiltered versions are far more robust in the dynamic conditions.

As anticipated, the standard-input optic flow algorithms produce nearly perfect results in static lighting conditions with a mean error of 1.6 pixels for both algorithms and a standard deviation of 0.77 and 0.79 for Horn–Schunck and Lucas–Kanade, respectively. The prefiltered version results overshot the mark slightly with mean errors of 4.8 and 5.9 pixels and standard deviations of 2.8 and 3.5 pixels for Horn–Schunck and Lucas–Kanade, respectively.

However, in dynamic lighting conditions the prefiltered results remain very close to the ground truth and do not deviate significantly over time. In this case, they averaged 1.8 and 4.1 pixels of error with standard deviations of 0.7 and 2.2 pixels for Horn–Schunck and Lucas–Kanade, respectively. Conversely, the standard input algorithms are not reliable under such conditions with mean errors of 10.9 and 31.45 pixels and standard deviations of 5.1 and 33.7 pixels for Horn–Schunck and Lucas–Kanade, respectively. The prefiltered optic flow does better in dynamic lighting than it does in static, and we are currently investigating how this can be generally applied.

To begin testing the boundaries of this method, we created the experiment presented in Fig. 5 by comparing a reference image at a known illumination with a series of other images of the same object captured at a variety of known illumination levels. The algorithms used here are identical to those used earlier in the report. For context, a change of 100 lux is within the deviation allowed within a well-lit office. Realistically, changes in illumination are often far more radical. A cloud passing in front of the sun can easily case variations on the order of 5000 lux. While the degree of variation in illuminance from frame to frame is dependent on the frame rate of the camera and the conditions the system is operating in, our method’s robustness to changes of more than 600 lux per transition, and our planned frame rate of 30 fps, give us confidence that even and indoor/outdoor transition of 10,000 lux would be manageable.

5.1.6 Conclusion

The primary assumption of traditional optic flow is that illumination remains constant, and under such conditions it performs admirably. However, the real world is a dynamic and irregular place, and any system that intends to operate within it must be robust to constantly changing conditions. As shown in Figs. 39, 40, and 41, simply by changing the input from a standard image to the double derivative of that image, we are able to produce a significant increase in position estimation accuracy under dynamic lighting conditions. The next big push in robotics is for small, autonomous systems to quickly navigate complex environments, specifically below the canopy in forests.[172]

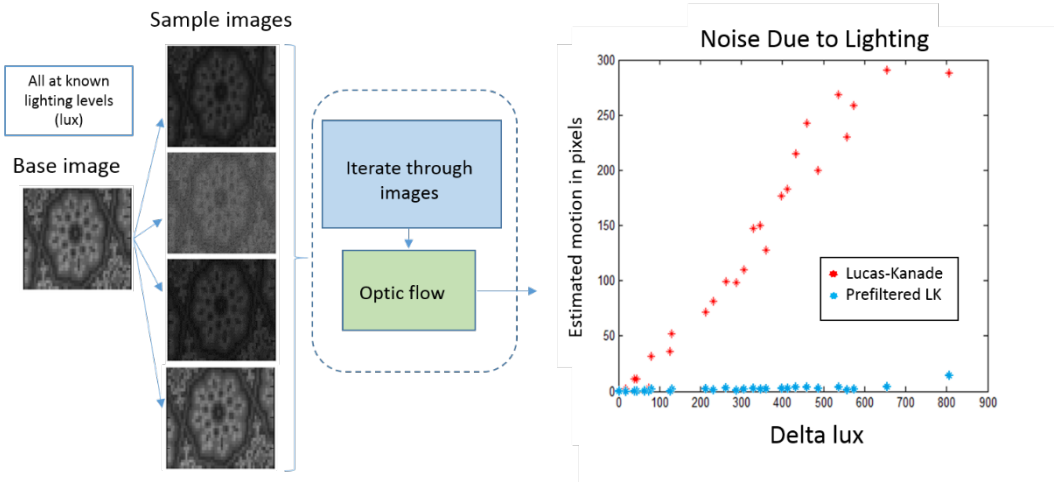

Fig. 41 Filtered and unfiltered optic flow's responses to extreme variations in illuminance

We have found that prefiltering is a useful tool for real-world navigation and particularly for size, weight, and power (SWaP)-constrained systems that cannot carry more than a single camera. It is computationally reasonable, as creating an image double derivate is substantially less expensive than the host optical flow algorithm, and it does not require any sort of a priori information. Our approach is robust to significant variations in lighting while also being very easily implemented and added to existing system. Despite the information lost by taking the derivative of the image prior to optic flow calculations, and the subsequent masking of the dead zones, we find that sufficient, relevant information persists for this method produces useful results.

5.1.7 Future Work

There is still a great deal of work to be done to make optic flow robust to vagaries of outdoor operation. We plan to investigate how best to identify and eliminate features due to strong shadows rather than actual environmental features. Despite

the strong edges and image variations caused by shadows, the textures of the objects within the scene remain the same. We have considered a method that compares the patterns and significant deviations.

Another typical source of error is dynamic backgrounds, such as windblown trees and grass. Such objects have fixed locations but nonetheless add a great deal of noise to the vector field and disrupt the flow patterns usually used to determine ego-motion and the presence of outside agents. Furthermore, when resident on a vehicle, there are likely to be ego-motion errors, such as vehicle drift and vibration that will need to be accounted for.

We look forward to combining all of these problems in future investigations into dynamic environments with highly constrained systems.

5.2 Dynamic Scenes

The present work was motivated by the authors' experiences during time spent embedded in an infantry unit in a forest in Columbus, Georgia.[171] We found that many elegant perception and localization algorithms fail dramatically when confronted with the realities of fielded operation. Forests are geometrically irregular and highly complex, consisting of heterogeneous and highly deformable elements, each of which may vary considerably from frame to frame in full-motion video. Additionally, there occur significant frame-to-frame perturbations in lighting and occlusion that have a significant effect on CV algorithms, the vast majority of which either explicitly or implicitly make rigid body assumptions about foreground objects and similarly rigid, regular lattice assumptions about background. We propose a method of computationally efficient object classification in a dynamic scene, where "dynamic" may refer to physical motion of the platform, background elements, and lighting occlusions. Certain phenomena, such as trees waving or people walking, have repeatable and recognizable patterns when viewed through an optic-flow vector field. In this report, we identify such patterns and show how to differentiate between dynamic and static background elements, mobile agents, and ego-motion produced by the platform. We emphasize that the purpose of this approach is not the explicit identification of people or trees but to differentiate between types of coherent motion and determine which areas are benign (e.g., trees having a limited and static area of influence) and which may constitute active threats, such as humans.

We will also discuss patterns of motion associated with other varieties of mobile agents, and show how these phenomena may be used to improve visual odometry based state estimation, by removing elements of the flow field that are not due to ego-motion.

The security field has been conducting significant investigations in detecting anomalies in crowds, as represented by a continuous flow field derived from video representations of pedestrians. Although this work is intended for security cameras, and as such is conducted from a static platform, it has many similarities to the initial stages of our work. Ryan et al.[202] identify disturbances from bicyclists and motor vehicles moving through areas that are typically populated only by pedestrians. As with our work, the focus is not on individual object detection but on general characterization of coherent motion patterns. However, the proposed disturbance detection method was found to be too computationally expensive for an aerial vehicle that must navigate in real time. Mehran et al.[203] also looked at the flow patterns created by large groups of people walking together and uses variations to identify anomalies. Unlike Ryan et al.,[202] whose computational burden came from overlapping regions of interest for multiple levels of examination, Mehran et al.[204] requires the full video sequence be ingested and processed forensically in batch mode. While this is acceptable for a static platform intended for surveillance and a posteriori analysis, it will not do for real-time flight.

Investigations of optic-flow-based obstacle avoidance from a mobile platform include the work in wide field integration by researchers at the University of Maryland.[204,205] Their work assumes an expected pattern of the unobstructed vector flow field and compares it with the input field. Disparities between actual and expected responses trigger evasive action, and they have been able to successfully navigate across cluttered environments. While this is an excellent system for known ego-motion and static backgrounds, it is unlikely that it would be able to handle a fully dynamic environment in which there were perpetual disparities between the responses.

Bideau and Learned-Miller[206] use optic-flow vectors to segment moving objects from complex backgrounds as observed from mobile platforms. However, their processing chain requires, among other steps, a random sample consensus evaluation of the estimated background motion. While this certainly assists in producing clean segmentation, sample consensus methods in general are too computational expensive to be responsive for our application space.

Our contribution is 2-fold. The first is an understanding of the level of danger presented by any particular phenomena without having to perform an in-depth analysis of the scene. From a navigational and collision avoidance perspective, the precise number of people in a region is irrelevant; however, knowing whether or not they are there or if the region is empty and all anomalies are due to wind-blown vegetation is extremely valuable. The second is that this approach is computationally lean. Rather than comparing large numbers of visual reference

features, we extract information from patterns in the vector field. A more thorough comparison is laid out in Section 5.2.2 of this report.

5.2.1 Data Set

We evaluated our method using data collected locally. We have chosen not to use either the KITTI[207] or Sintel[201] data sets. While captured on a mobile platform, KITTI largely observes unpopulated, static environments. Sintel is animation, and, as such, the motion of the characters and scenic elements do not directly correlate to their real-world counterparts. We prefer not to introduce unnecessary noise into our problem space.

Many of the publically available datasets, including those previously discussed, are not representative of Army-relevant conditions; as such, we have chosen to collect our own. We are in the process of publicly releasing it in the hope that it will prove useful to the community. Our videos were first collected from a tripod to test motion in the environment, exclusive of platform noise, and then from onboard a DJI Inspire One quadrotor and a handheld GoPro camera to add in noise due to ego motion, platform jitter, and the like. All data were collected outdoors, in the mid-Atlantic United States in fair weather. Videos include a variety of backgrounds, wind conditions, standoff distances from targets of interest, image resolutions, and an assortment of mobile-agent target types. These mobile agents include a variety of people and gait styles, other robots, and motor vehicles. Video sequences were captured at approximately 30 Hz; however, these have been downsampled to closer to 10 Hz to enhance the visibility of ego motion between frames.

In future versions of this work we intend to include clips from the Berkeley Segmentation data set,[208] as well as other relevant publically available sets, to allow for true comparison with state-of-the-art systems.

5.2.2 Technical Approach

Our method is predicated on the assumption that a precalculated optic flow vector field has been produced by a vision-aided state estimation algorithm, such as Parallel Tracking and Mapping[209] and other visual odometry techniques in an IMU-fusion framework,[210,211] and is readily available for our use. By utilizing precomputed flow fields, we can reduce the overall computational requirements of the system. A natural limitation of this method is that it will only work in environmental conditions conducive to optic flow; total darkness or featureless environments are known sources of failure for any optical-flow-based technique using visible-spectrum EO data. By considering the motion vectors in terms of polar

coordinates, they are easily divisible into magnitude and orientation. As seen in Fig. 42, different phenomena present different-magnitude signatures. Static scene elements will show very little local variation in the magnitude of its vectors. Dynamic scene elements, such as mobile agents or windblown vegetation, will show significant local variation. During our time embedded with an infantry unit, the primary sources of scene dynamics were humans and vegetation. Differentiating between these 2 patterns and ego-motion was the first phase of our experiment, and we find it is still a useful dichotomy to explain the methodology. The forms taken by these local variations very with the phenomena from which they are produced. Regardless of the degree of the oscillation experienced by a tree in the wind, all the branches are moving together, and there is limited local variation. Conversely, a walking human will present a large variety of vector magnitudes representing the separate motion of the arms, legs, head, and torso.

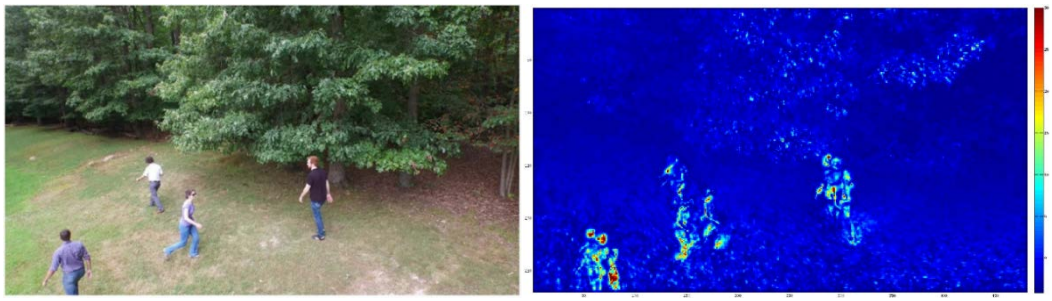

Fig. 42 People walking in front of windblown trees as seen through (left) one of the original images used in the optic-flow calculation and (right) the vector magnitudes it produced. The range in magnitudes is described by the colors: dark red being the longest and dark blue the shortest.

We have experimentally determined that the disturbance patterns are regular, reproducible, and agonistic to the physical characteristics of the system, such as lens size and shape or imager size and resolution. The magnitudes of the vectors vary with the capture modality, capture rate, and platform velocity, but the patterns remain consistent. Although the examples shown are limited to vegetation and humans (Fig. 43), the methodology is easily expandable to other categories. Motor vehicles, an oft-requested category, appear as a group of spatially related vectors with identical orientations and minimal variations in magnitude.

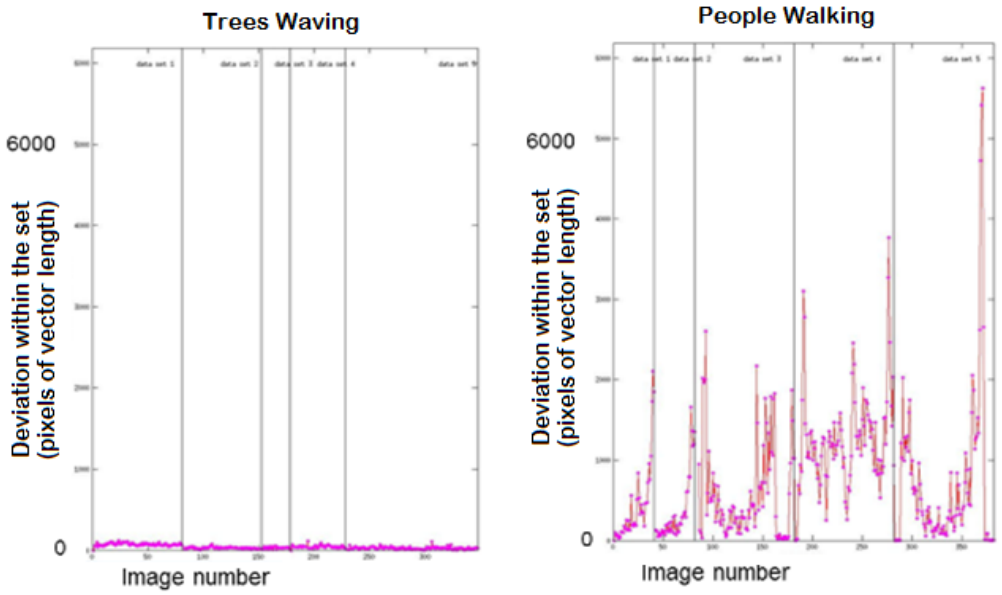

Fig. 43 Variation in magnitude of optic flow vectors of 2 types of dynamic scene elements, as represented by 5 video sequences for each

5.2.3 Methods

The categorization methodology has been intentionally kept as lean as possible, to ensure its fieldability on a constrained processors resident on man-packable UAS.

To investigate local variations, the optic flow field of an image pair is first divided up into sections (known as "bins") in an *nxm* array, and the variance of the vector magnitudes is calculated for each (Eq. 27), where *G* is the variance field, *x* and *y* are pixel locations in the original image, μ is the mean over the grid cell, $n \in N$, and $m \in M$.

$$G(n,m) \triangleq \sum_{x',y'} \frac{OF(x',y') - \mu_{n,m}}{\frac{X}{N} * \frac{Y}{M}},$$

where

$$x' \in \left[n\frac{X}{N}, (n+1)\frac{X}{N} \right] \quad (27)$$

and

$$y' \in [m\frac{Y}{M}, (m+1)\frac{Y}{M}].$$

As ego-motion will add additional "noise" to the classification process, we estimate its impact and remove it from the values to be considered. This is accomplished by determining the maximum variance value over all bins (Eq. 28) and determining which bins in the set are less than or equal to the ego-motion scale factor multiplied by the maximum value (Eqs. 29 and 30). In the case of the data presented in this paper, that scale factor is 0.7, which was arrived at empirically. Where *B* is the

maximum over all the grid cells, Γ defines the bins to be included, and T is a threshold, currently set to 0.7,

$$B = \max_{\substack{n\in N \\ m\in M}}(G(n,m)). \tag{28}$$

$$\Gamma(n,m) \triangleq \delta_{nm} * G(n,m), \tag{29}$$

where

$$\delta_{nm} \triangleq \begin{cases} 1 \; if \; G(n,m) < T * B \\ 0 \; otherwise \end{cases}. \tag{30}$$

Ego-motion is estimated to be the mean variance of those bins that fit the preceding criteria (Eq. 31), where P is the estimated ego-motion.

$$P = \frac{\sum_{n,m}\Gamma(n,m)}{\sum_{n,m}\delta_{nm}}. \tag{31}$$

The ego-motion estimation (mean variance) is subtracted from all the original variance values in each of the bins so as to bring the magnitudes in line with what would be observed by a stationary platform (Eq. 32). Where H is estimated scene motion,

$$H(n,m) \triangleq G(n,m) - P. \tag{32}$$

Categorization thresholds are applied to determine how the degree of variation found in each of the bins compares with previously calculated thresholds (*T1-6*) (Eq. 33).

$$J(n,m) = \begin{cases} \mathrm{T1} < \mathrm{H(n,m)} < \mathrm{T2} \\ T3 < H(n,m) < T4. \\ T5 < H(n,m) < T6 \end{cases} \tag{33}$$

When capturing imagery on a static platform, it is possible to use hard-coded thresholds for agent classification. However, when working with mobile platforms, the vectors created by the dynamic scene elements are added to those created by the ego-motion of the vehicle, which renders hardcoded values obsolete. Nevertheless, the disturbance patterns remain consistent regardless of the scale of the vectors. To ensure that the thresholds match the scale of the image, we apply a scale factor based on image size. The initial threshold values are derived from the vector magnitudes of the base image size. The base size was arbitrarily chosen from the data collected, as some sort of standard comparison point was required. The current image resolution is compared with that of the base image, and the correct fraction is multiplied against the classification variance thresholds.

Our system requires no a priori information about the image resolution, and recalculates the thresholds and ego-motion on the fly. The threshold values cannot

be hardcoded for 2 reasons. The first is that doing so would limit the ease of deployment on novel platforms. The less we have to change, the better. The second is the increasing reliance on algorithms that would change the resolution of imagery data depending on mission need. As such, we anticipate having to be robust to significant variation even within a single mission profile.

We are fully aware that this system does not offer perfect accuracy; however, exact segmentation is not particularly relevant for our application. Our objective is not to correctly classify all potentially relevant segments but rather to classify enough to keep from crashing. We offer further discussion of this point in Section 5.2.4.

What this approach offers is computational efficiency, which is particularly suited to small-scale UASs. As the Army moves toward man-packable, cargo-pocket UASs, we anticipate severely SWaP and computationally limited platforms that are not capable of significant processing loads. When an image is divided into *nxm* bins, *C* is the overhead to load the images and calculated averages, and *P* is the number of pixels per bin, the computational load is described as

$$Load = (4NMP) + C. \tag{34}$$

In the case of a 12 × 12 array of bins, this works out to be approximately 576*P+16 computations per image, regardless of image size. The processors being used by our team (Odroid-XU4) are capable of 4000 million instructions per second, which leaves the overwhelming majority of the processor available for other requirements and goals.

5.2.4 Results

A video of the algorithms accurately finding mobile agents, in a variety of scenes from a variety of platforms may be found at https://github.com /usarmyresearchlab/dynamic scenesvideo.

Unlike traditional CV investigations, we are less interested in precisely segmented obstacles than in a functional fieldable system. It does not really matter if there are 3 people standing in a clump or 5, or even just one. What is important is the ability to mark coarse segments of the frame as obstacle locations or locations that are safe for travel, as well as the ability to determine which static scene elements may be used for platform state estimation (Fig. 44).

Fig. 44 Do we really care there are 3 people or 4—or, just that some areas are clear and others are not? It really does depend on the application.

To quantitatively measure results in the current stage of development, we are looking at accuracy as compared with ground truth, mean processing time per image set, and FP rate. To count as a true FP, we required a 40% overlap of each annotated area with the classification regions (Fig. 45).

Fig. 45 (left) Classifications (red boxes) overlaid with annotations (cyan lines) and (right) true and FPs and negatives

5.2.5 Conclusion

In addition to being useful for scene categorization, the framework presented in this work would be a useful step toward improving visual state estimation. If all of the identified areas of "noise" in the flow field can be removed, the remaining flow vectors may be determined to represent the ego motion of the platform.

Although the scope of this report is focused on the feasibility of using optic flow vectors as a scene categorization technique, in future work we will investigate the utility of this method to identify anomalies in the vector field and the degree of improvement in state estimation when mobile agents are removed and only ego-motion is utilized.

This methodology is robust to a variety of realistic environmental conditions, image resolutions, and platform dynamics (Table 14). While it does not purport to find every disturbance in the flow field, it finds them with sufficient frequency to add value to a state estimation system. Static objects in the scene will either not move at all, as seen from a static platform, or will move together, as seen from a mobile platform. As such, this approach is not intended for individual object detection and

classification. Rather, it is intended to eventually help eliminate noise sources from state estimation problems and as assist in coarse obstacle avoidance.

As described later, the US Army Maneuver Center for Excellence's (MCOE's) Infantry School has shown interest in this methodology and has requested that we test it under a variety of scenarios. Additionally, this method could be used for efficient postprocessing of the enormous quantity of data being produced by national assets. Due to the massive proliferation of sensors and data-acquisition platforms fielded, the majority of collected data are never reviewed, and human operators spend an inordinate amount of time on data that are not "interesting" from a mission perspective. If an algorithm could highlight areas of interest, human operators would be able to focus their time and attention on relevant video clips, with potential positive effects on operator efficacy and fatigue.

Table 14 Video clips and associated metrics

Video clip	Image resolution	Annotated agents in video clip	Agents detected in video clip	Mean percentage FPs	Average time to process frame (s)
3 people in woods	540×960×3	57	39	5	0.1
One person in front of trees	1296×2304×3	17	17	3	0.2
People walking away on path	650×1201×3	122	87	7	0.1
2 people walking near trees	1296×2304×3	66	66	7	0.2
People in woods, off the path	601×1200×3	127	64	8	0.1
People running near 507	1080×1920×3	201	130	7	0.1
From above, at 507	270×480×3	126	74	2	0.09
People throwing leaves	540×960×3	100	46	9	0.1
3 cars driving	324×576×3	20	19	4	0.09
Truck driving	1296×2304×3	8	5	5	0.2
3 people on path	1080×1920×3	23	21	17	0.1
People and a quadrotor	501×1001×3	89	70	16	0.1

6. Dynamic Teaming

In pursuit of autonomous systems that are able to operate effectively in dynamic environments with human teammates, we focused on developing online and distributed techniques for general machine learning. Online techniques perform machine learning incrementally as data are provided, giving an agent the ability to adapt to new and changing information in real time. Distributed techniques perform a joint machine-learning task across a network, enabling a team of agents to jointly learn from the experiences of individual agents and thereby reducing the overall time it takes for the team to adapt to new information. Our main result is Decentralized Dynamic Discriminative Dictionary Learning (D4L): an algorithm we developed to perform online, distributed, discriminative dictionary learning and applied to a task involving several unmanned ground vehicles (UGVs) exploring a new environment. To broaden the applicability, we extended the concepts behind D4L in 2 ways: 1) by adapting the framework such that it can be used for a lower-level, self-supervised robot-maneuver task and 2) by developing a more general online technique capable of performing nonparametric function approximation. D4L has been featured in several publications.[212–217]

6.1 Decentralized Dynamic Discriminative Dictionary Learning

D4L is an algorithm for discriminative dictionary learning in a distributed online setting.[1–3] D4L provides a new framework with which a team of networked autonomous agents may jointly perform dictionary learning using the observations of the constituent members. We applied D4L to the problem of a robotic team seeking to autonomously classify textures in an unknown environment.

Dictionary learning is an unsupervised technique for creating a low-dimensional representation of a data domain that is constructed from the observed data rather than relying on a hand-engineered representation. In practice, having such a low-dimensional representation of the input domain is critical to developing tractable models for regression or classification. With D4L, we simultaneously learn the representation (dictionary) as well as the classification model on top of this representation, also known as Discriminative Dictionary Learning.

Because the algorithm is distributed, we can perform this joint representation and classifier learning in domains where different agents have access to different portions of the data distribution (e.g., a team of ground robots occupying disparate regions of a battlefield [Fig. 46]). In such a case, D4L can effectively use the team's communication network to transfer knowledge gleaned from the observations of one agent to the others, even when the other agents have not made similar

observations themselves. And because D4L is an online algorithm, a team that uses it is able to continuously update the learned knowledge and therefore adapt to any changes in the environment.

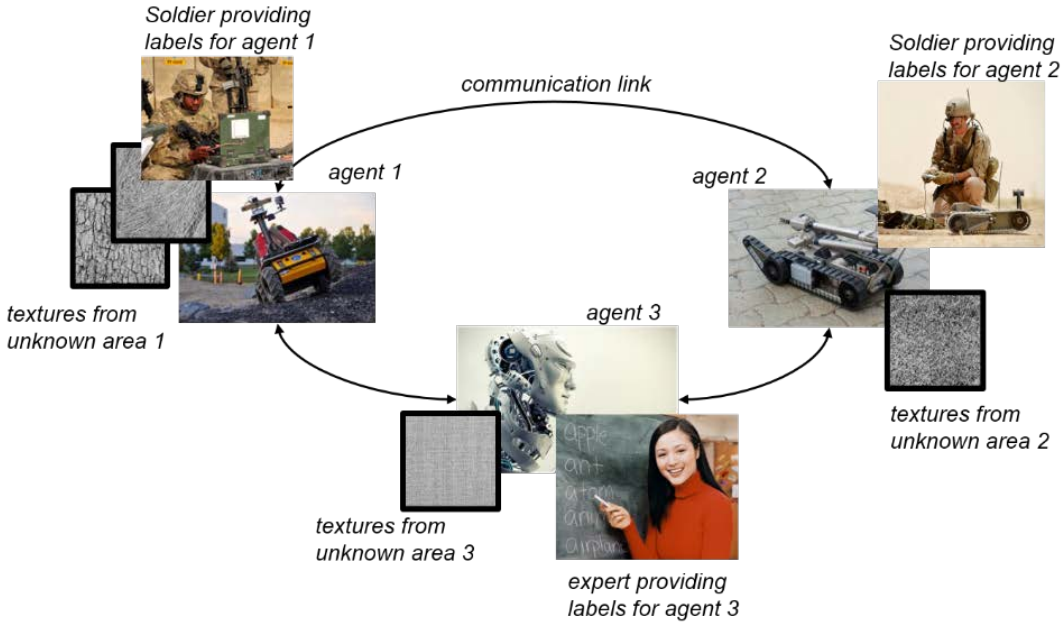

Fig. 46 Sample task for which D4L may be applied: Individual autonomous agents aim to jointly learn how to classify textures using their own observations, information from human teammates, and model information transmitted over the network.

In our work we specifically considered the case in which a network of agents aims to estimate model parameters, including a common set of feature-space dictionary elements, for a supervised machine learning task on the basis of sequentially received observations. We formulated this problem as a distributed stochastic program with a nonconvex objective and proposed using a block variant of the Arrow–Hurwicz saddle point algorithm to solve it. Model information need only be exchanged between neighboring nodes, and we enforced consensus through Lagrange multipliers that penalize the model discrepancy between neighbors. Our theoretical results demonstrated that decisions made with this saddle point algorithm asymptotically achieve a first-order stationarity condition on average for a learning rate that depends on the specific signal source, network, and discriminative task.

Our experiments focused on learning to classify textures such as the one seen in Fig. 47. Such texture classification is a useful precursor for evaluating navigability in driving tasks and could form the basis for the navigation system of a UGV.

Fig. 47 Sample texture from the Brodatz texture database

For certain network structures, we observed that the performance of the distributed D4L algorithm was comparable to that of a centralized version (Fig. 48). We also analyzed the learning capability of D4L when individual agents made observations of their own unique part of the data space. That is, each agent was restricted to an "incomplete" view of the data space while, overall, the team makes observations from the overall distribution. While we observed slightly worse performance in this case (compared with one in which each agent made observations of the whole space), we also note that D4L did allow the network to jointly improve its performance as more observations were made (Fig. 49).

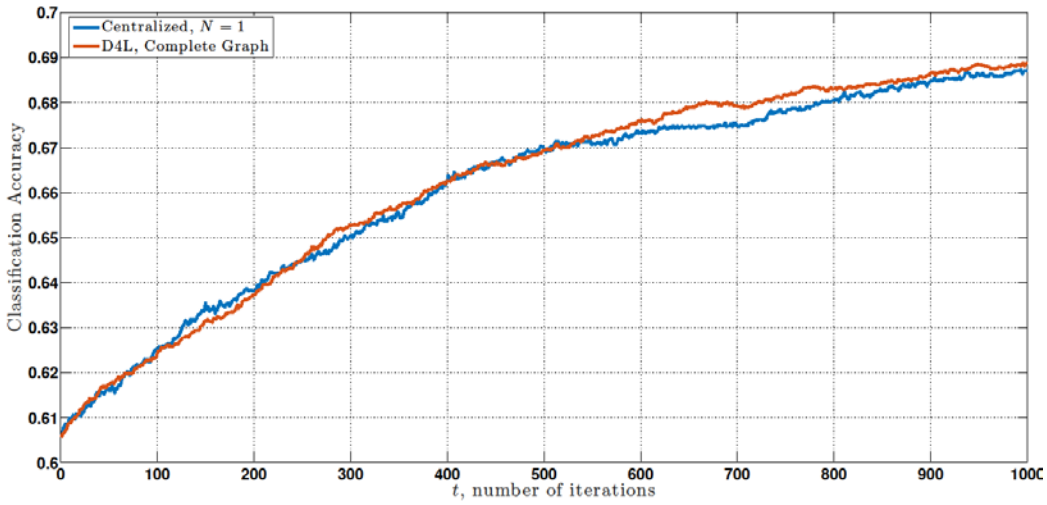

Fig. 48 Performance of D4L for the centralized and complete-graph scenarios: For this network structure, the distributed algorithm performs just as well as its centralized counterpart.

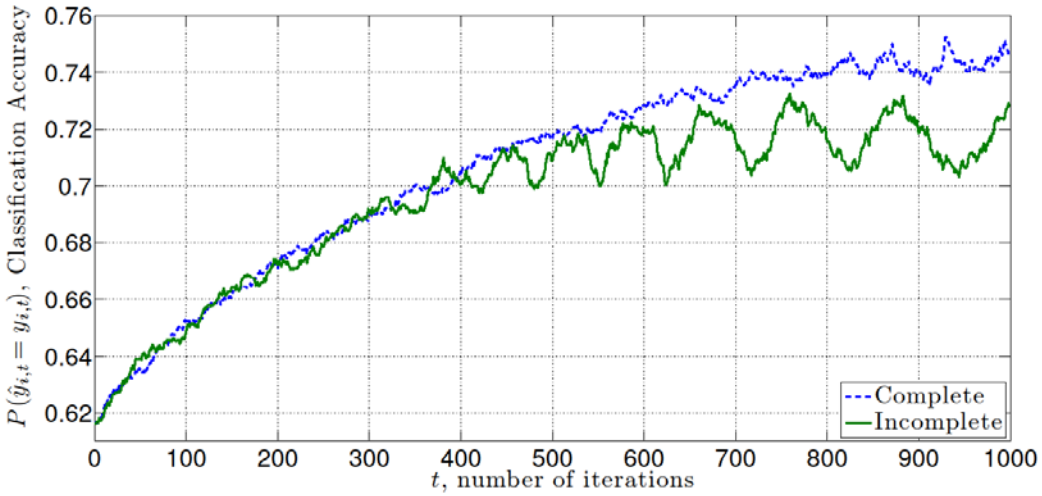

Fig. 49 Performance of D4L for complete and incomplete data observations: "Complete" refers to case in which each agent made observations over entire data space; "Incomplete" refers to case in which each agent made observations from its own unique part of the data space.

Finally, we also analyzed the case in which the network size varies. As one might expect, joint learning in the network happened more slowly with an increase in the number of nodes. However, joint learning occurred in all cases we considered (Fig. 50).

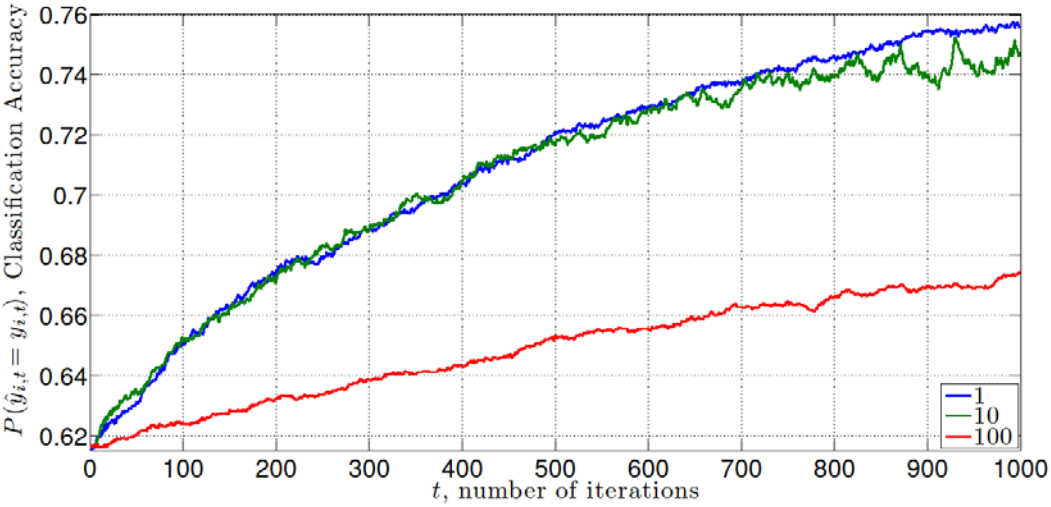

Fig. 50 Performance of D4L for networks with a varying number of nodes

6.2 Online Learning for Characterizing Unknown Environments in Ground Robotic Vehicle Models

Inspired by the success of D4L, but motivated by the desire for a less-human-intensive task, we considered the problem of trying to increase the autonomous driving performance of a UGV in unmodeled environments (Fig 51), a learning task that is more self-supervised in nature. We sought to predict the distribution of structural state-estimation error due to poorly modeled platform dynamics as well as environmental effects. Such predictions are a critical component of any modern control approach that uses uncertainty information to provide robustness in control design. We used an online-learning, algorithm-based D4L to fit a statistical model of error that provides enough expressive power to enable prediction directly from motion control signals and low-level visual features.[4] We called the proposed technique Online Learning for Drivability Assessment (OLDA).

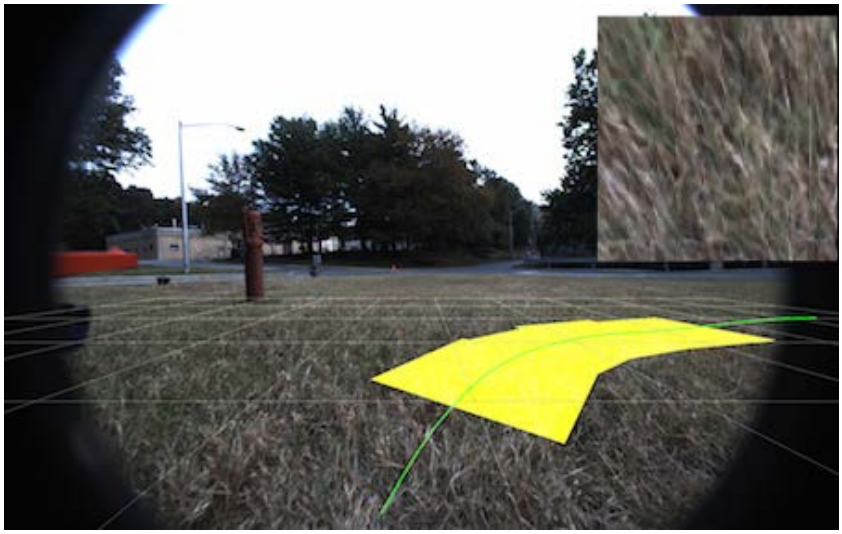

Fig. 51 Robot-centric view of the environment: In this work, we develop a way to predict driving model disturbance on the basis of the camera image and the planned path, as shown here.

OLDA models disturbances as Gaussian random variables, where the mean and variance are each functions of both visual features of the local terrain and the desired control signal. Since the relationship between the model disturbance and these quantities is likely to be extremely complicated, we expected it to be highly nonlinear and therefore modeled the mean and variance as functions of sparse codes of the data observations. We formulated the problem of learning the mean and variance functions as a task-driven dictionary learning problem (as in Section 6.1) and proceeded to use a centralized learning algorithm to perform the function estimation.

We quantified the advantages of our algorithm in a real-world experimental setting. To do so, we collected data on an iRobot Packbot, which is a ground platform equipped with a skid-steer tracked drive system with onboard computation and, among other sensors, an IMU to detect disturbance and a camera to gather images of the terrain. In Fig. 52, we compare our OLDA technique with that of 2 others: 1) an average model that estimates the disturbance as the running average over all past disturbance data and 2) a windowed average model that performs the averaging only over a recent time period. While Fig. 52 shows our dictionary-based model achieved lower loss values than these other models, the real effect of this is seen in Fig. 53, where the predicted distribution in blue seems to match the actual disturbance data quite well. In Fig. 54, the effect of this on positional state forecasting can be seen in that our technique provided predictions that characterize well the actual path traversed by the platform.

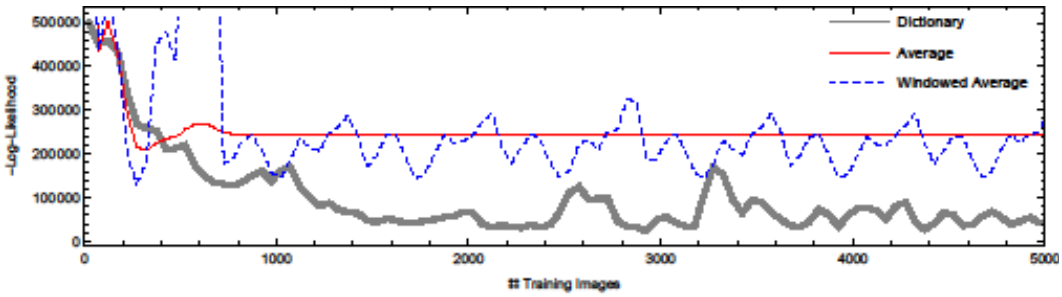

Fig. 52 Loss values for our model (gray) and the average (red) and windowed average (blue) techniques—lower is better

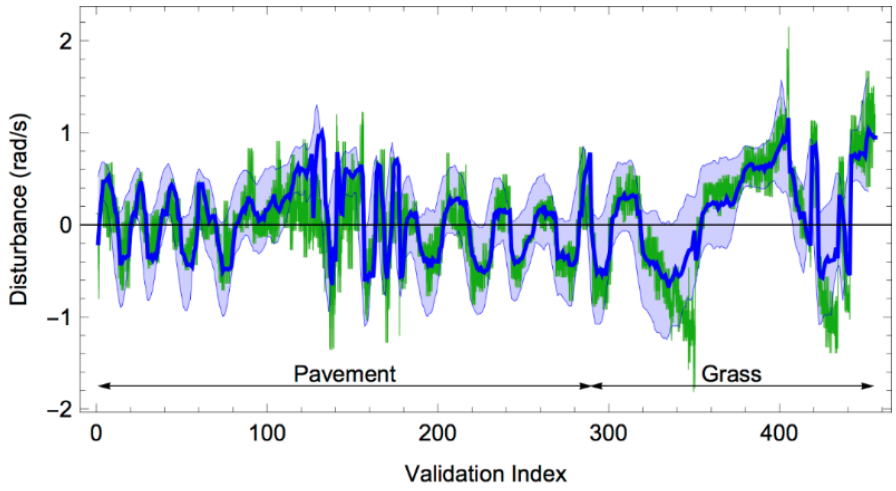

Fig. 53 Statistics of the disturbance prediction across test set is visualized in blue as a solid line for the mean predicted disturbance and a shaded envelope depicting the "two-sigma" envelope; true disturbance is shown in green.

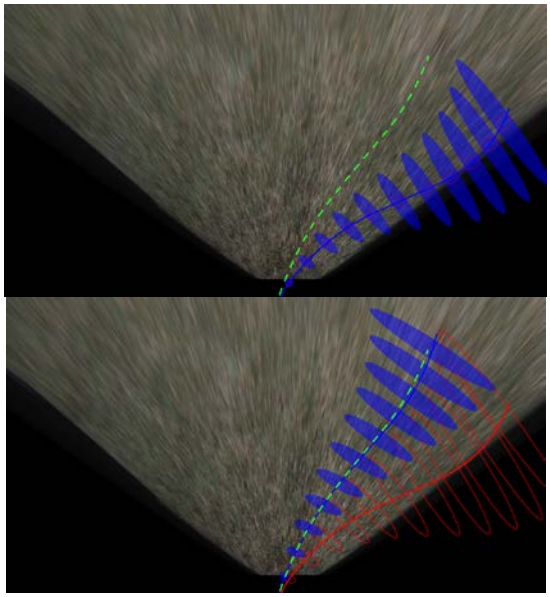

Fig. 54 State uncertainty propagated according to model prediction and control-input time series for an example drawn from the terrain and grass test sets before training (top) and after training (bottom). Green dashed line is actual driven path; blue-filled ellipses show prediction based on our dictionary learning algorithm; red path/ellipses depict the average model. Prediction generated by our method almost exactly matches actual disturbance experienced by the platform, meaning we successfully predicted where steering mistakes were likely along a future reference trajectory.

6.3 Parsimonious Online Learning with Kernels via Sparse Projections in Function Space

Given the reasonable success of parametric methods in an online setting, we also worked on extending nonparametric techniques to be applicable in an online setting. Despite their attractiveness, popular perception is that techniques for nonparametric function approximation do not scale to streaming data due to an intractable growth in the amount of storage they require. To solve this problem in a memory-affordable way, we proposed an online technique based on functional stochastic gradient descent in tandem with supervised sparsification based on greedy function subspace projections. The method, called Parsimonious Online Learning with Kernels (POLK),[215.216] provides a controllable tradeoff between its solution accuracy and the amount of memory it requires.

More specifically, POLK is a new technique for learning nonparametric function approximations in a Reproducing Kernel Hilbert Space that respects optimality and ameliorates the complexity issues classically associated with such techniques. We accomplished this by 1) shifting the goal from that of finding an approximation that is optimal to that of finding an approximation that is optimal within a class of parsimonious (sparse) kernel representations and 2) designing a training method

that follows a trajectory of intermediate representations that are also parsimonious. The algorithm also admits theoretical guarantees that neither complexity nor (lack of overall) optimality become untenable.

We validated POLK using several datasets, including a synthetic one drawn from a multiclass Gaussian mixture model (see Fig. 55), the Mixed National Institute of Standards and Technology (MNIST) handwritten digits (Fig. 56), and the texture database we used when evaluating D4L. For each data set, we saw that POLK compared favorably with other online nonparametric techniques (Figs. 57 and 58), and found that it was the only algorithm that could be used for several data set/loss combinations (e.g., logistic regression on MNIST). Moreover, compared with batch solutions like that of the popular LIBSVM software, POLK yielded a test-set error just 1.0% higher while using an order of magnitude fewer model points. Additionally, POLK is able to run online, with streaming data, whereas batch solutions like LIBSVM cannot.

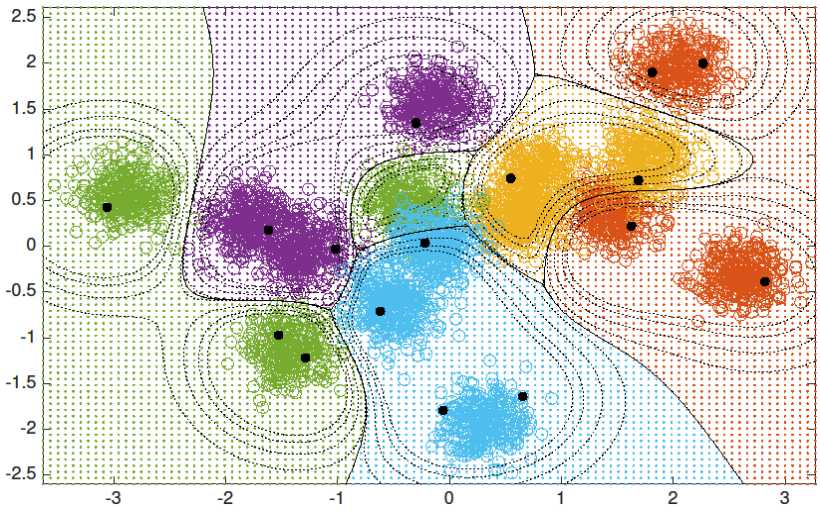

Fig. 55 Synthetic data set and learned kernel logistic regressor: Training examples from distinct classes are assigned a unique color. Grid colors represent the classification decision of the learned classifier; bold black dots are selected kernel dictionary elements concentrating at modes of the joint data distribution; solid curved lines show the class boundaries, while the dashed depict the confidence intervals.

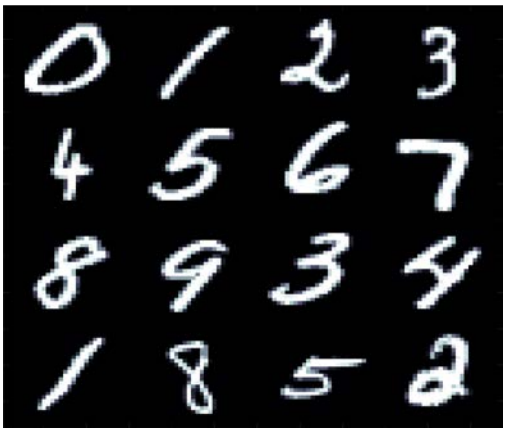

Fig. 56 Example images from the MNIST handwritten digit data set

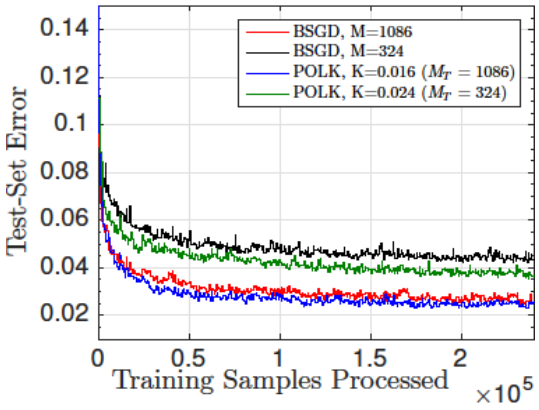

Fig. 57 Classification error for the MNIST data set using the hinge loss function

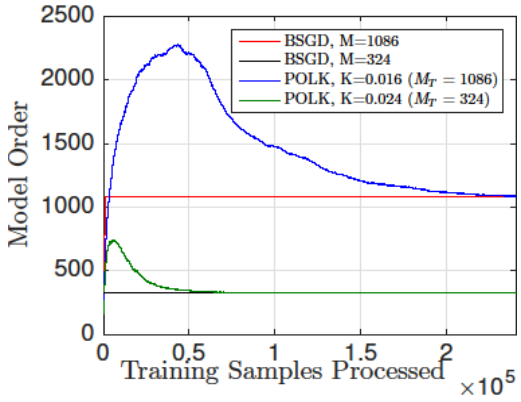

Fig. 58 Model order for MNIST data set using the hinge loss function; POLK has a model order that is able to change during the learning procedure

7. Integrated Experiment

To close out this 3-year effort, we have begun an integrated experiment in which a heterogeneous network of human and robot agents are tasked to identify potential targets and explore a novel environment in real time. Our initial efforts have been directed toward the target-identification task. Here, we describe our progress examining the effect of variable communications bandwidth and connectivity on the collective target-identification performance of this heterogeneous team. Subsequently, we will outline the next steps toward an integrated simulation of the effects of intermittent or variable communications on simultaneous target identification and exploration by a heterogeneous network.

7.1 Target Identification

The target-identification task examines the effect of constrained communications on target-detection performance. We define this scenario as shown in Fig. 59. An autonomous agent is engaged in target detection and exploration and acquires high-resolution images from its environment. Using its onboard resources, the autonomous agent is capable of performing some limited CV analysis on the images in an attempt to identify targets of interest. Additionally, the agent can downsample or compress the images to transmit them back to the tactical operations center (TOC). The level of compression can be tailored to account for the communication constraints at that time. At the TOC, images sent from the autonomous agent are processed by both human and CV agents. At the TOC, the CV agents are free of many of the processing restrictions that may exist on a fielded autonomous agent (memory, power, etc.); thus, the CV agents are able to make use of state-of-the-art approaches to identify targets of interest. Labels from both the CV agent and human agent are then fused and relayed back to the autonomous agent where the image originated. The decision from the TOC may be combined with a decision made locally by the autonomous agent.

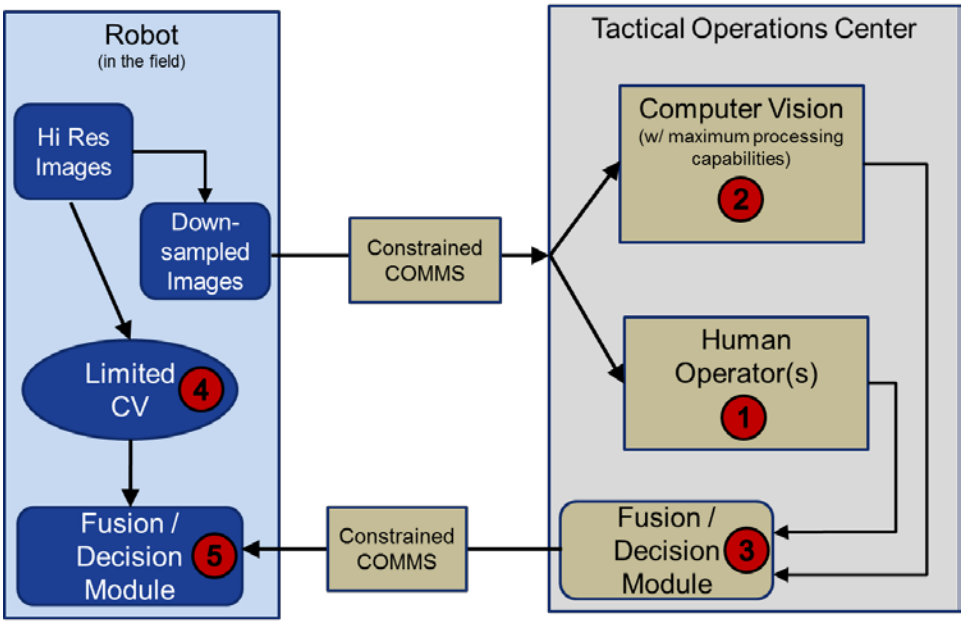

Fig. 59 Schematic of the target-identification scenario used within the integrated experiment

In this scenario, we imagine time to be of the essence. Under the assumption of unlimited processing power at the TOC, CV algorithms can be assumed to return an immediate result, but human image analysts might respond slowly, especially when given an open-ended search task. RSVP, in which human agents are shown a series of images in rapid succession (here, 4/s) with the task of pressing a button whenever a target image is seen, can dramatically increase the throughput of examined images without substantially degrading image-labelling performance.[39,40]

Within this scenario, there are a number of questions that warrant further exploration (numbered red circles in Fig. 60).

1) How does human target-identification performance change when viewing compressed images? Can we develop appropriate confidence measures to capture the variability in performance in these situations?

2) How does CV-based target-identification performance change when using compressed images? We can also explore variants of this question that look at the effect of training and testing the CV algorithms on either the same or different levels of compression. Large drops in performance when compression levels between training and testing sets are mismatched indicate that systems would require pretraining at prescribed compression levels to make such a situation plausible. Conversely, the lack of significant drops in performance when compression levels between training and test

sets are mismatched would indicate that a generically trained classifier might work across a range of compression values.

3) How is the fusion of human and CV decisions impacted when attempting to identify targets in compressed images? Does the utility of each agent vary with compression level?

4) Which CV methods can be deployed onto an autonomous platform? Such a method would need to minimize power and memory consumption and use limited processing power to apply CV algorithms to a novel image. We should also differentiate between algorithms that can be trained on the autonomous agents versus algorithms that would need to be trained in a more traditional computing environment. This differentiation becomes important if we need to adapt the models of targets during operations. Algorithms that can be efficiently trained without extensive processor, memory, or power requirements are candidates for adaptive systems.

5) How does the severity of degraded communications impact the fusion of a limited CV algorithm operating on full-resolution images and joint human–CV analysis on compressed images? Under what conditions does the limited CV algorithm enhance the decisions made by the joint analysis of compressed images? Under what conditions does the joint analysis of compressed images fail to improve the decisions provided by the limited CV algorithms?

In the efforts completed thus far, we have run an experiment that will allow us to answer Questions 1–3. The remaining 2 questions will continue to be examined in the near future. Here, we describe the methods we used in collecting the data for our experiment along with plans for data analysis.

7.1.1 Impact of Image Compression on Human Target Identification Performance

7.1.1.1 Participants

Participants (N = 16, all right-handed, mean age = 31.2, standard deviation = 7.8) had normal or corrected-to-normal vision and were free of neurological illness or trauma by self-report. The voluntary, fully informed written consent of participants in this research was obtained as required by federal and US Army regulations.[112,113] The investigator adhered to Army policies for the protection of human subjects.[113] All human subjects testing was approved by the ARL's Institutional Review Board.

7.1.1.2 Stimuli

Stimuli for this experiment came from a library of digital photographs (512 pixels wide by 662 pixels tall, encoded as lossless jpeg files) taken in and around a large mixed-use office complex. A more complete description of this data set can be found in Touryan et al.[130]; relevant details are included here. This complex included occupied and unoccupied office space, laboratory space, classroom space, and outdoor areas. In contrast with many existing image libraries in which some object of interest is centered in the frame, these photos were taken without any particular object or framing intention. No formal procedure for randomizing the photographs was used at the time of photography, but the intent was to simulate a random sampling of views from a cluttered office-like environment that an autonomous system might encounter when exploring such an environment.

The images in this library were manually tagged with metadata, including whether the images included each of 5 object categories (chairs, doors, stairs, containers, and posters). From this library of 3000 images, 1800 were selected for use in this experiment. A total of 180 images that included chairs were pseudorandomly selected as target stimuli, and 1,620 nontarget images were pseudorandomly selected from the remaining images that did not include chairs. Prior to selection, all images were reviewed and those that included objects that might be semantically ambiguous chairs (e.g., stools, benches, and couches) were excluded from selection.

A main factor of interest in this experiment was image compression. To achieve 4 distinct levels of compression, the JPEG2000 compression algorithm[218] as implemented in MATLAB version 2014a was used. This algorithm was chosen for its good performance, convenience of use and because the standard implements several features that make it especially suitable for wireless transmission in bandwidth-limited circumstances, including bit-error resilience and transmission at increasing levels of detail. Requested compression ratios, in terms of the ratio of number of bits in the uncompressed image (i.e., 24 bits per pixel) to the number of bits in the compressed image, were 1 (i.e., uncompressed), 500, 1000, and 2000. The implementation of the JPEG2000 compression algorithm we used does not afford precise control over compression levels, so actual compression ratios were measured from the resulting files. For the requested ratio of 500, actual compression ratios ranged from 500.2 to 584.7 (mean = 506.1, median = 503.6). For the requested ratio of 1000, actual compression ratios ranged from 1000.8 to 1290.4 (mean = 1018.2, median = 1011.8). For the requested ratio of 2000, actual compression ratios ranged from 2001.6 to 2661.9 (mean = 2051.0, median = 2033.7). Example images at the 4 compression levels appear in Fig. 60.

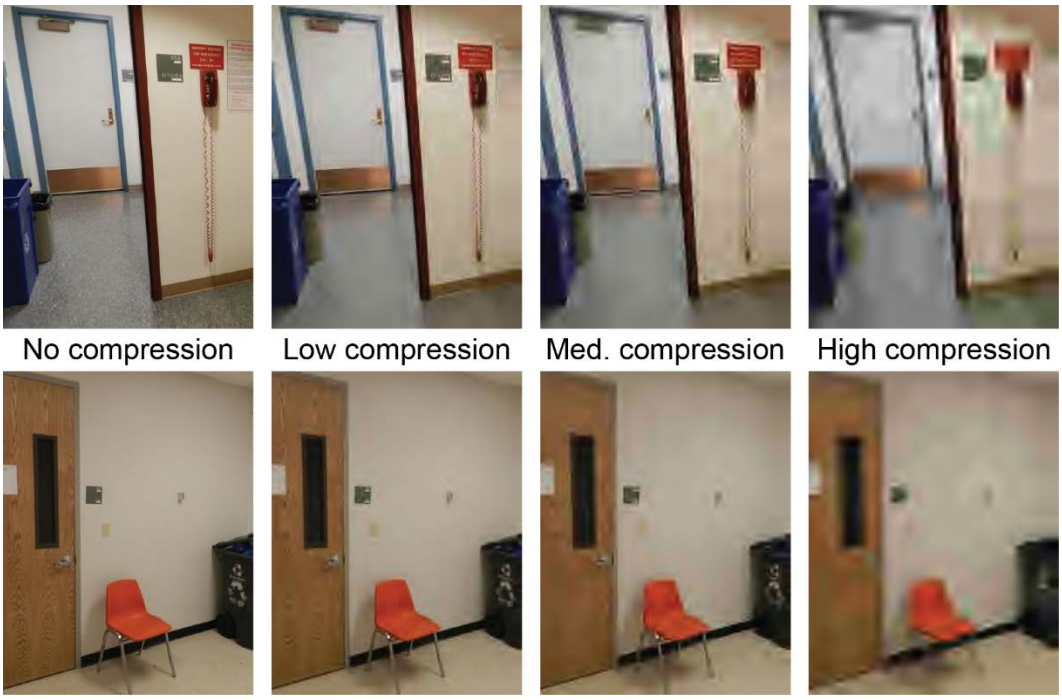

Fig. 60 Example images at the 4 levels of compression: no compression, low (500:1), medium (1000:1), and high (2000:1)

7.1.1.3 Procedure

Images were displayed to the participant on a 24-inch, liquid-crystal display monitor at a distance of approximately 0.7 m. Images were centered and subtended approximately 11.4° by 14.7° of visual angle. Images were displayed in a rapid serial visual-presentation paradigm, at a rate of 4 images/s. After every 9 s of stimulus presentation, a screen with the word "blink" appeared for 2 s to allow participants time to blink without missing a stimulus presentation.

Participants were instructed to monitor the stream of images and press a button on a response box whenever they saw a chair in the image. Of the 1800 images in the experiment, 180 included chairs, and the rest did not include any chair. This resulted in a target frequency of 0.1.

The full experiment was divided into 5 blocks. The first block was a familiarization block in which 180 images were shown to the participant who was then offered the opportunity to ask clarifying questions about the task. Data collected during the familiarization block were not analyzed. After any questions were answered, the remaining 4 experimental blocks were run. At the midpoint of each block and between each block, the participant was given the opportunity to take a self-paced break.

Each block consisted of all 1800 images presented at a particular compression level. The order of compression level was counterbalanced across subjects. Within each block, stimulus order was pseudorandomized with constraints. To facilitate cross-validation, the 1800 images were each assigned to belong to one of 10 equally sized chunks. The order of the chunks within a block was pseudorandomized for each block. The effect of this chunking procedure was to ensure that the images within a chunk were always displayed in a contiguous period of time. An additional constraint was on the placement of targets in the image stream. Target frequency was 0.1 within each chunk, and images were ordered such that at least 2 nontarget images were interposed between target images.

7.1.1.4 Behavioral Analysis

BPs were collected using a Cedrus (San Pedro, California) response box, which offers 2- to 3-ms response-time precision. From these responses, participant performance, expressed as HR and FAR, and a classification score were calculated using the method of Files and Marathe,[219] which will be summarized here. For each subject, a sample of response times was obtained by looking for any responses that occurred within 1.5 s of a target image throughout that subject's run. From this sample, an RT-PDF was estimated as an ex-Gaussian distribution with parameters set using maximum likelihood estimation.[51] This RT-PDF was used to apportion credit for each BP response to the preceding stimuli according to their relative likelihoods. The expected value of this apportionment function for stimulus at time S_i is given in Eq. 35 (Files and Marathe[219], Eq. 6):

$$E[A(S_i)] = HR \times \sum_{S_j \in tar} \sum_{T} \left[f(S_j - T) A(S_i, T) \right] + FAR \times \sum_{S_j \in n.t.} \sum_{T} \left[f(S_j - T) A(S_i, T) \right] z \,, \qquad (35)$$

where f is the RT-PDF, T is the time the response occurred, S_j is the time at which other recent stimuli occurred, and $A(S, T)$ is the apportionment onto a stimulus at time S of a response at time T. Because all quantities on the right-hand side of the equation are known except HR and FAR, this equation reduces to $HR \times A_i + FAR \times B_i$, for each stimulus i. The values of HR and FAR are solved by regression against the actual attribution each image received, yielding an estimate of HR and FAR for that block.

For BP-based classification, each compression-level block was treated separately. Within a block, cross-validation was done with respect to chunks of 180 images, all of which were displayed in one contiguous period of time. For each of 10

cross-validation folds, 6 chunks were selected as training chunks, and the remaining 4 were considered testing chunks. Using only the training chunks, an RT-PDF and subject performance (HR and FAR) were estimated using the previously described method. Also using only the training chunks, each image in the training chunk was assigned a score using the method described. Next, a threshold was found that minimized a hinge loss function with the slope of the loss function defined to pass through the threshold at zero and to reach 1 at the extreme (minimum or maximum) of all the scores in the training set. To summarize, from training data we estimated the subject's HR and FAR for that block, the subject's RT-PDF, a minimum and maximum score for the images in the training block, as well as an optimized threshold for separating targets from nontargets.

For testing, images were assigned a score based on the apportionment of each BP onto the preceding stimuli according to their relative likelihoods from the trained RT-PDF. As an estimated confidence on the score, the hinge loss function from training was used to assign confidence equal to the loss expected if this label is incorrect. When combining a BP-derived score with other classifier modalities, it is weighted by a confidence measure derived from the block-level performance of that participant estimated from the training set as p(Target)*HR/(p(Target)*HR + (1-p(Target))*FAR). This is the prior probability that a response from this participant is in response to a target image.

7.1.1.5 Electroencephalography Recording and Preprocessing

Data from 64-channel EEG were recorded using a Biosemi ActiveTwo system. In addition, 6 external input channels were used to acquire bipolar EOG measurements related to horizontal and vertical eye movements and mastoids. EEG signals were digitized at 2 kHz and offline were decimated to 512 Hz.

Offline, EEG data were imported into EEGLAB,[220] rereferenced to the average of the mastoid electrodes, and band-pass filtered to 0.1–50 Hz using the EEGLAB function pop_eegfiltnew. To remove the potential influence of eye blinks on the EEG, an independent components analysis was run using Infomax ICA,[221] and those components with strong correlation with the EOG were manually inspected and eliminated from the data if they appeared to reflect genuine eye movement or eye blink activity.

7.1.1.6 Hierarchical Discriminant Component Analysis

HDCA is a single-trial classification method that can be used to determine the presence of an ERP related to the target within the EEG signal. HDCA is a binary classification method based on an ensemble of logistic regression classifiers.

HDCA transforms multichannel EEG data collected over a temporal window relative to image onset into a single interest score. Ideally, the interest score is generated so that the range of scores for each class are distinct, thereby allowing for simple discrimination of the 2 classes.

Generating interest scores from HDCA involves a 2-stage classification. In the first stage, a set of 10 logistic regression discriminators are applied to 10 equally sized, nonoverlapping time windows that range from image onset up to 1-s postimage onset. The choice of 10 Stage-1 discriminators was based on previous studies.[60] Other studies have used 20 Stage-1 discriminators and reported no significant difference when comparing classification performance using 10 versus 20 Stage-1 discriminators. Each of the 10 discriminators are trained independently. Each of these 20 discriminators serve to collapse the information contained in all 64 EEG channels collected over the course of the corresponding time window into a single value for discriminating between the neural signals evoked by the 2 image classes. In the second stage, a separate logistic regression discriminator is applied to the output of the 10 Stage-1 discriminators to create a single interest score that can efficiently discriminate between the 2 image classes.

Similar to the BP analysis, for HDCA classification each compression-level block is treated separately. Within a block, cross-validation is done with respect to the chunks of 180 images, all of which are displayed in one contiguous period of time. For each of 10 cross-validation folds, 6 chunks are selected as training chunks, 2 chunks are selected as testing chunks, and the final 2 chunks are left out for validation of fusion across multiple classifiers. Fusion analyses are described in more detail in the following.

7.1.1.7 Fusion of Neural Classifier and Behavior

The purpose of including a NC is to assess whether it provides any additional information to the button score, particularly at high image-compression levels when behavioral performance degraded. The added value of the NC is measured by comparing the accuracy of target identification of either classifier alone against the accuracy when both HDCA and behavioral result are combined using a fusion classifier such as the dynamic belief fusion method described in Section 4.

7.1.2 Impact of Image Compression on CV-Based Target-Identification Performance

Having described the method by which our data were collected and the plans for analysis of both behavioral and neural data to further understand how image degradation affects human target detection performance, we now turn to how such degradation might affect CV algorithms.

7.1.2.1 Overview of CV Analyses

The CV analyses are designed to complement the analyses applied to the human data described above. As such, CV-based object detection algorithms are applied to the same sets of images that were presented to the human participants.

7.1.2.1.1 CV Algorithms and Sensor Fusion

The individual images are each processed by a set of algorithms to classify target images. The selected algorithms use different feature extraction methods (e.g., HOG, dense SIFT, color attributes, and CNN features) and different principles of detecting objects of interest. This heterogeneity in detectors enhances the possibility that each algorithm will contribute unique information regarding each image. Each algorithm, when applied to an image, produces a classification score that indicates a degree of confidence about a particular decision (target versus nontarget). The following specific algorithms are applied:

- DPM: This method represents objects as a set of parts that can be deformed using 2 different scales of HOG features, latent features, and a deformation cost.[222]
- SVM-based Dense SIFT: This method is based on matching densely sampled, pixelwise SIFT features between 2 images while preserving spatial discontinuities.[139]
- ESVM: This method learns a separate classifier for each positive training image, using a rigid HOG template, and scores candidate detections based on "distance" to exemplars.[125]
- CNNs: This method leverages a deep learning framework to combine feature extraction and classification on images. Two variants are explored here.[142, 223]

The output of each of the specified CV algorithms will be fused to generate an overall CV classification for each image. The output of this fused CV classifier will be used to address Question 2 (Section 7.1 and Fig. 60). The output of the overall CV classification will be fused with the output of the classifiers operating on the human data to generate an overall human–CV classification for each image. The output of the overall human–CV classifier will be used to address Question 3 (Section 7.1 and Fig. 60).

Both fusion stages (CV and human–CV), will apply the DBF approach described in Section 4.[118] Briefly, DBF is a novel late-fusion framework that models joint relationships between a priori and current information of individual detectors. This approach robustly extracts complementary information from multiple information sources.

7.1.2.1.2 Cross Validation

To enable direct comparison and fusion of CV results with the human results, the training and testing of the individual object detection algorithms and the fusion algorithms follow a similar approach to cross-validation as the human data described previously. At all stages, the classifiers and fusion algorithms are trained and tested on non-overlapping data. Table 15 details the iterative 6–2–1–1 cross-validation sequence used. Table 15, Line A, shows that the first 6 of the 10 total chunks provide training data for the individual classifiers. Chunks 7 and 8 are then used as both test data for the individual classifiers and as training data for the CV fusion classifier that combines data across the individual object detectors. Chunk 9 provides the test data for the CV fusion classifier and training data for the human–CV fusion classifier, which combines human and computer outputs. Finally, Chunk 10 provides the test data for the final result of the human–CV classifier. This procedure is then repeated, such that in the second iteration Chunks 2–7 train the individual classifiers, Chunks 8 and 9 test the individual classifiers and train the CV fusion classifier, and Chunk 10 trains the human–CV classifier, which is then tested on Chunk 1 (Table 15, Line B). This 6–2–1–1 cross-validation sequence is carried out 10 times (Table 15, Lines A–J), rotating the data subsets such that, when complete, each image category serves as both training and test data for each level of integration.

Table 15 Cross-validation procedure for training and testing human and CV classifiers

Decision level	1	2	3	4
Train	**Individual classifiers**	**CV fusion classifier**	**Human–CV fusion classifier**	**…**
Test	**…**	**Individual classifiers**	**CV fusion classifier**	**Human–CV fusion classifier**
A	1–6	7–8	9	10
B	2–7	8–9	10	1
C	3–8	9–10	1	2
D	4–9	10,1	2	3
E	5–10	1–2	3	4
F	6–10,1	2–3	4	5
G	7–10,1–2	3–4	5	6
H	8–10,1–3	4–5	6	7
I	9–10, 1–4	5–6	7	8
J	10, 1–5	6–7	8	9

7.1.2.2 Summary and Impact

The approach described here will enable an effective fusion of human and CV target-identification decisions. As such, this work should clarify the impact of degraded communications on a heterogeneous human–autonomous-agent team's overall target-identification performance. However, when autonomous agents are tasked with finding targets in an unknown environment, 2 tasks must be accomplished simultaneously: target identification and environment exploration. As a future direction of this work, we describe a scenario on which simulations can be based to see the impacts of degraded communications on team performance.

7.2 Scenario Overview

We describe a scenario upon which simulations may be built in which a mixed team of human and robot agents is tasked to explore an unknown environment and identify targets of interest. To facilitate experimentation, and to enable examination

of the broadest set of questions, both the exploration and target identification will take place in the simulated environment represented by the MOUT-site database described in Section 3 and Fig. 61.

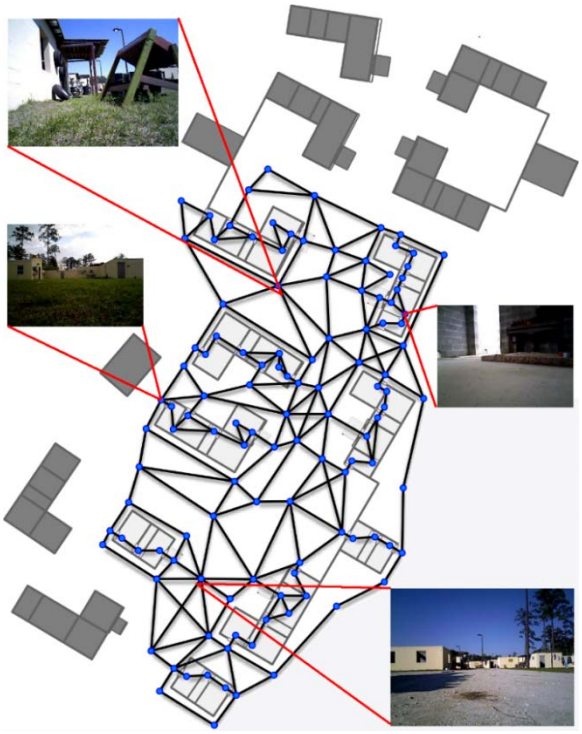

Fig. 61 Schematic and example views from the MOUT site database: nodes (blue dots) are both indoor and outdoor locations; connecting paths (black lines) show where the robot travelled; selected frames illustrate differences in lighting conditions and scene types.

The database includes video clips used in this experiment that were extracted from a data set of robotic sensor readings at a MOUT training site. This data set was captured to simulate an exhaustive exploration of the site, and some example views are in Fig. 3. Nodes representing a physical location and pose (facing north, south, east, or west) were defined both indoors and out, and nearby nodes, including all at the same location with different poses, were connected to each other by paths. An iRobot Packbot, equipped with a Prosilica high-resolution color camera (resolution 2752 × 2200, frame rate 1 Hz), an Asus Xtion Pro RGB-D camera (resolution 320 × 240, frame rate 30 Hz), an actuated Hokuyo LiDAR, and a Garmin GPS), traversed these paths at a speed of about 1.2 m/s while recording all sensor data.

Orange soccer balls were included at various points throughout the environment as potential targets of interest. While their inclusion was meant to facilitate incorporation of some type of target-identification tasks within the environment, the brightly colored soccer balls tend to easily stand out. As such, both human and CV agents are easily able to identify the targets without much error. Because of the ease with which both agents could identify these targets, we use the soccer balls as

target markers rather than targets themselves. As target markers, these soccer balls represent locations where the autonomous agent and the human simulate a series of joint target-identification tasks under specific communication constraints using a separate target-identification data set (as described in Section 7.1).

7.2.1 Planned Simulations

The integrated simulations are designed to examine the impact of team composition and communications quality on the overall performance of a mixed-agent team on 2 different tasks. Each simulation run involves a team of one human agent and 2–4 autonomous agents. They are tasked with exploring a novel environment that is instantiated by using a subset of the MOUT-site database with targets located in up to 10 randomized locations. The autonomous agents are randomly placed within the environment, while the human agent is assumed to be operating from a nearby TOC location. The autonomous agents begin exploring the environment and generating a map. For the exploration, we will compare a variety of established techniques with some of the novel approaches we developed in this project. As the exploration proceeds, each time an autonomous agent encounters the target marker (i.e., orange soccer ball), a target-identification session is triggered, sampled from the target-identification experiment already done. Once the target-identification session is complete, the results from that session are saved, and the autonomous agents continue exploration. As the robot explores the environment, if a potential loop closure is identified a human user is queried to determine if the loop closure does in fact exist. This query is resolved by looking at the results of loop-closure experiment described in Section 3. These processes continue until the autonomous agents have fully mapped the location and have achieved a prescribed level of performance on each of the target-marker locations. For each run, the time to completion, accuracy of the map, and accuracy of target detection will be used to quantify performance. This entire process can be repeated with different numbers of agents, different numbers of target-marker locations, and various levels of communications constraints.

This simulation framework will allow us to build upon already completed work to more fully explore the impacts of information variability in heterogeneous human–autonomous teams carrying out a simultaneous exploration and target-identification task.

8. Work Products and Transitions

We have divided our work products and transitions into sections, with summary statistics listed in parenthesis when appropriate. Over 3 years, we have supported 5 postdoctoral fellows and one contractor with onsite research projects. Of these individuals, 2 have been hired as civilian employees and one has transitioned into civilian service with the US Navy at the Space and Naval Warfare Systems Command in San Diego, California. Additionally, one of our researchers was hired directly into civilian service. We have also identified several transition points for the research conducted under this DSI project.

The human-variability research focused on developing confidence measures that estimate expected performance of human agents on the basis of behavioral and physiological data. This research will continue under fiscal year 2017 (FY17) Big Idea as Continuous Multi-faceted Soldier Characterization for Adaptive Technologies. This project will focus on developing a basic understanding of the factors that contribute to variability in human behavior and techniques to predict changes in performance. Furthermore, the techniques developed here will also be used by a new DSI entitled Decentralized Learning of Network-of-Networks of Heterogeneous Multi-Agent Systems.

The CV and sensor fusion work focuses on developing novel CV and sensor fusion methods. The sensor fusion approaches have been shown to extract complementary information from multiple sources, and may be useful for applications beyond the scope of CV alone. This work will continue under the FY17 project Vision Aided Position, Navigation, and Timing (PNT) in Contested Environments. Additionally, transitions are being explored with Communications-Electronics Research, Development and Engineering Center/NightVision.

The agent-adaptation research focused on creating sensing techniques that are robust to complex and dynamic environments to enable operational-tempo navigation in real-world scenarios. This work will continue under the FY17 project Vision Aided PNT in Contested Environments. The MCOE's Infantry School has shown interest in this methodology and has requested that we test it under a variety of scenarios, to include forests, around the motor pool, amid cluttered urban areas, in a barracks with people, around other robots, and on and off road. They have requested an algorithm that is sufficiently reliable for the safety teams to allow Soldiers within 2 m of the robots. Solders are now prohibited from coming within 12 m of UGVs, as their obstacle-avoidance capabilities are insufficiently reliable.

Work on dynamic teaming has planned transitions into 3 major efforts. First, the FY19 Big Idea, Novel Forms of Joint Human–System Decision Making, has a focus

on data- and time-constrained deep learning and learning with statistically mismatched data. Our work on nonparametric techniques (POLK) provides an important alternative to traditional parametric deep-learning techniques because it is possible to geometrically characterize where there are gaps in the model due to missing data, and we plan to use this work as a starting point for alternative research under this upcoming program. Second, the FY17 DSI project, Modeling Networks of Networks of Heterogeneous Multi-Agent Teams, focuses on adapting reinforcement learning using human interaction. The reference architectures for this task are built upon deep learning using convolutional neural networks, but we anticipate that our work on nonparametric techniques (POLK) could provide an alternative formulation for some or all of the architecture and provide an intuitive way to incorporate examples demonstrated by a human. Third, we are further expanding the OLDA work as part of a reinforcement learning framework to develop control policies for robot navigation in outdoor environments as part of our FY17 Information Sciences Division mission proposal.

We have 4 published journal publications: one in press, one under review, and 2 in preparation for submission. We also have 13 published conference proceedings and one in press, 2 under review, and 3 in preparation. We also have one published ARL technical reports and 2 more in preparation for submission.

8.1 Personnel

We have built an in-house research group through several mechanisms to enhance ARL's in-house capability to perform research related to teaming of heterogeneous agents. Five postdocs have been involved in onsite projects. Three of those 5 postdocs transitioned to civilian employees (Amar Marathe, Jamie Lukos, and Benjamin Files). Dr Amar Marathe assumed the role of principal investigator of this DSI shortly after transitioning. Dr Jamie Lukos joined as a civilian employee at the Space and Naval Warfare Systems Command in San Diego, where she continues to interact with ARL researchers. Dr Benjamin Files joined ARL–West and is working to build up research efforts in that facility. Garrett Warnell joined ARL as a civilian employee early in this DSI effort. He has since transitioned from Adelphi to ARL–South. Allison Mathis joined ARL as a contractor and continues to work with the Sensors and Electron Devices Directorate.

8.2 Journal Publications (4 Published, 1 in Press, 1 under Review, 2 in Preparation)

1. Files BT, Lawhern VJ, Ries AJ, Marathe AR. A permutation test for unbalanced paired comparisons of global field power. Brain Topography. 2016;29(3):345–357.
2. Files BT, Marathe AR. A regression method for estimating performance in a rapid serial visual presentation target detection task. Journal of Neuroscience Methods. 2016;258(30):114–123.
3. Marathe AR, Ries AJ, Lawhern VJ, Lance BJ, Touryan J, McDowell K, Cecotti H. The effect of target and nontarget similarity on neural classification performance: a boost from confidence. Frontiers in Neuroscience. 2015;9:270.
4. Tsiligkaridis T, Sadler BM, Hero AO 3rd. On decentralized estimation with active queries. IEEE Transactions on Signal Processing. 63.10. 2015:2610–2622.
5. Koppel A, Warnell G, Stump E, Ribeiro A. D4L: decentralized, dynamic, dictionary learning. IEEE Transactions on Signal and Information Processing; 2017 in press.
6. Koppel A, Warnell G, Stump E, Ribeiro A. Parsimonious online learning with kernels via sparse projections in function space. Journal of Machine Learning Research; 2017 submitted.
7. Lee H, Kwon H, Robinson R, Nothwang W, Marathe AR. Dynamic belief fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence; 2017 in preparation.
8. Lee H, Kwon H, Robinson R, Nothwang W, Marathe AR. Human-machine sensor fusion in unified tasks. IEEE Transactions on Systems, Man, and Cybernetics; 2017 in preparation.

8.3 Conference Publications (13 Published, 1 in Press, 2 under Review, 3 in Preparation)

1. McDowell K, Marathe AR, Lance BJ, Metcalfe JS, Sajda P. Neuro-robotic technologies and social interactions. Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction; 2014.
2. Koppel A, Warnell G, Stump E, Ribeiro A. D4L: decentralized dynamic discriminative dictionary learning. Presented at the 2015 IEEE/RSJ

International Conference on Intelligent Robots and Systems; 2015 Sep 28–Oct 2; Hamburg, Germany.

3. Koppel A, Warnell G, Stump E, Ribeiro A. A stochastic primal-dual algorithm for task-driven dictionary learning in networks. Presented at the Asilmoar Conference on Signals, Systems, and Computers; 2015 Nov 8–11; Pacific Grove, CA.

4. Robinson R, Lee H, McCourt M, Marathe AR, Kwon H, Ton C, Nothwang W. Human-autonomy sensor fusion for rapid object detection. Presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems; 2015 Sep 28–Oct 2; Hamburg, Germany.

5. Lee H, Kwon H, Nothwang W, Robinson R, Marathe AR. Dynamic belief fusion for object detection. Proceedings of the Winter Conference on Applications of Computer Vision; 2016:1–9.

6. Lee H, Kwon H, Robinson R, Nothwang W. DTM: deformable template matching. Presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing; 2016 Mar 20–25; Shanghai, China.

7. Lee H, Kwon H, Robinson R, Donavanik D, Nothwang W, Marathe AR. Task-conversions for integrating human and machine perception in a unified task. Presented at the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems; 2015 Sep 28–Oct 2; Hamburg, Germany.

8. Files BT, Canady J, Warnell G, Stump EA, Nothwang WM, Marathe AR. Human assisted robot exploration. Presented at SPIE Defense and Commercial Sensing; 2016 Apr 17–21; Baltimore, MD.

9. Shamwell EJ, Lee H, Kwon H, Marathe AR, Lawhern VJ, Nothwang W. Single-trial EEG RSVP classification using convolutional neural networks. Presented at SPIE Defense and Commercial Sensing, 2016 Apr 17–21; Baltimore, MD.

10. Lee H, Kwon H, Robinson RM, Nothwang WD, Marathe AR. An efficient fusion approach for combining human and machine decisions. Presented at SPIE Defense and Commercial Sensing; 2016 Apr 17–21; Baltimore, MD.

11. Mathis AM, Donavanik D, Nothwang WD. Computationally Efficient scene categorization in complex dynamic environments. Presented at the IEEE Applied Imagery Pattern Recognition Workshop; 2016 Oct 18–20; Washington, DC.

12. Koppel A, Fink J, Warnell G, Stump E, Ribeiro A. Online learning for characterizing unknown environments in ground robotic vehicle models. Presented at the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems; 2015 Sep 28–Oct 2; Hamburg, Germany.

13. Bency A, Kwon H, Lee H, Vadivel K, Manjunath BS. Weakly supervised localization using deep feature maps. Presented at the European Conference on Computer Vision; 2016 Oct 8–16; Amsterdam, The Netherlands.

14. Koppel A, Warnell G, Stump E, Ribeiro A. Parsimonious online learning with kernels via sparse projections in function space. Proceedings of the IEEE International Conference on Audio, Speech, and Signal Processing; 2017 in press.

15. Cao Y*, Lee H*, Kwon H. Enhanced object detection via fusion with prior beliefs from image classification. IEEE International Conference on Image Processing (ICIP); 2017 submitted. (*indicates equal contribution.)

16. Eum S*, Lee H*, Kwon H, Doermann D. OI-CNN: integrating architecturally different object detection networks for event recognition. IEEE International Conference on Image Processing (ICIP); 2017 submitted. (*indicates equal contribution.)

17. Files BT, Canady JD, Drnec K, Marathe AR. Image compression degrades both behavioral and neural classification performance on RSVP task; 2017 in preparation.

18. Lee H, Kwon H, Nothwang W, Canady JD, Files BT, Marathe A. Perceptions on the compressed RSVP dataset; 2017 in preparation.

19. Lee H, Kwon H, Nothwang W, Canady JD, Drnec K, Files, BT, Marathe A. Late fusion of human and machine perceptions via DBF on the compressed RSVP dataset; 2017 in preparation.

8.4 Technical Reports (1 Published, 2 in Preparation)

1. Mathis A, Nothwang W, Donavanik D, Conroy J, Shamwell J, Robinson R. Making optic flow robust to dynamic lighting conditions for real-time operation. Adelphi Laboratory Center (MD): Army Research Laboratory (US); 2016 Mar. Report No.: ARL-TR-7629.

2. Canady J, Files BT, Marathe AR. An implementation of a regression-based method for estimating performance in a rapid serial visual presentation

target detection task. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2017 in preparation.

3. Canady J, Files BT, Marathe AR. Development of an interface for incorporating human-assistance to improve loop closure detection in robotic exploration. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2017 in preparation.

9. References

1. Christensen JC, Estepp JR, Wilson GF, Russell CA. The effects of day-to-day variability of physiological data on operator functional state classification. NeuroImage. 2012;59(1):57–63.

2. Liao LD, Lin CT, McDowell K, Wickenden AE, Gramann K, Jung TP. Biosensor technologies for augmented brain: computer interfaces in the next decades. Proc IEEE. 2012;100:1553–1566.

3. Cunningham A, Wurm KM, Burgard W, Dellaert F. Fully distributed scalable smoothing and mapping with robust multi-robot data association. Proceedings of the 2012 IEEE International Conference on Robotics and Automation (ICRA). 2012;53:1093–1100.

4. Olson E, Strom J, Goeddel R, Morton R, Ranganathan P, Richardson A. Exploration and mapping with autonomous robot teams. Commun ACM. 2013;56(3):62–70.

5. Stroupe AW, Martin MC, Balch T. Distributed sensor fusion for object position estimation by multi-robot systems. Proceedings of the IEEE International Conference on Robotics and Automation. 2001;2:1092–1098.

6. Scerri P, Pynadath D, Tambe M. Adjustable autonomy for the real world. In: Agent autonomy. Berlin (Germany): Springer; 2003. p. 211–241.

7. Archer S, Warwick W, Oster A. Current efforts to model human decision making in a military enviornment. Simul Ser. 32(3):151–155.

8. Janssen M, Ostrom E. Empirically based, agent-based models. Ecol Soc. 11(2):37.

9. Martino BD, Kumaran D, Seymour B, Dolan RJ. Frames, biases, and rational decision-making in the human brain. Science. 2006;313(5787):684–687.

10. Weiss G. Multiagent systems: a modern approach to distributed artificial intelligence. Cambridge (MA): MIT Press; 1999.

11. Sycara K, Pannu A, Willamson M, Zeng D, Decker K. Distributed intelligent agents. IEEE Expert. 1996;11(6):36–46.

12. Luo X, Jennings NR, Shadbolt N, Leung H, Lee JH. A fuzzy constraint based model for bilateral, multi-issue negotiations in semi-competitive environments. Artif Intell. 2003;148(1–2):53–102.

13. Manyika J, Durrant-Whyte H. Data fusion and sensor management: a decentralized information-theoretic approach. Upper Saddle River (NJ): Prentice Hall; 1995.

14. Gonzalez C, Vanyukov P, Martin MK. The use of microworlds to study dynamic decision making. Comput Hum Behav. 2005;21(2):273–286.

15. Newman P, Leonard J, Tardos JD, Neira J. Explore and return: experimental validation of real-time concurrent mapping and localization. Proceedings of the IEEE International Conference on Robotics and Automation. 2002;2:1802–1809.

16. Jensen FV, Nielsen TD. Bayesian networks and decision graphs. Berlin (Germany): Springer; 2007.

17. Dias MB, Zlot R, Kalra N, Stentz A. Market-based multirobot coordination: a survey and analysis. Proc IEEE. 2006;94(7):1257–1270.

18. Tanner HG, Jadbabaie A, Pappas GJ. Flocking in fixed and switching networks. IEEE Trans Autom Control. 2007;52(5):863–868.

19. Ratcliff R, Philiastides MG, Sajda P. Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. Proc Natl Acad Sci. 2009;106(16):6539–6544.

20. Sheridan TB. Function allocation: algorithm, alchemy or apostasy? Int J Hum-Comput Stud. 2000;52(2):203–216.

21. Dekker SW, Woods DD. MABA-MABA or abracadabra? Progress on human–automation co-ordination. Cogn Technol Work. 2002;4(4):240–244.

22. Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern Part Syst Hum. 2000;30(3):286–97.

23. Fitts PM. Human engineering for an effective air-navigation and traffic-control system. Fort Belvoir (VA): Defense Technical Information Center (US); 1951. File No.: ADB815893.

24. Sheridan TB. Telerobotics, automation, and human supervisory control. Cambridge (MA): MIT Press; 1992.

25. Crandall JW, Goodrich MA. Characterizing efficiency of human robot interaction: A case study of shared-control teleoperation. Proceedings of the Intelligent Robots and Systems IEEE/RSJ International Conference; 2002. p. 1290–1295.

26. Crandall JW, Goodrich MA. Experiments in adjustable autonomy. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics; 2001, p. 1624–1629.

27. Sellner B, Heger FW, Hiatt LM, Simmons R, Singh S. Coordinated multiagent teams and sliding autonomy for large-scale assembly. Proc IEEE. 2006;94(7):1425–1444.

28. Fong T, Thorpe C, Baur C. Multi-robot remote driving with collaborative control. IEEE Trans Ind Electron. 2003;50(4):699–704.

29. Fong T, Thorpe C, Baur C. Robot, asker of questions. Robot Auton Syst. 2003;42(3–4):235–243.

30. Bernoulli D. Exposition of a new theory on the measurement of risk. Econom J Econom Soc. 1954;22(1):23–36.

31. Pascal B, Krailsheimer AJ. Pensees. New York (NY): Penguin; 1968.

32. Lehmann EL. Some principles of the theory of testing hypotheses. Berlin (Germany): Springer; 2012.

33. Tsiligkaridis T, Sadler B, Hero A. Collaborative 20 questions for target localization. IEEE Trans Inf Theory. 2014;60(4):2233–2252.

34. Hayati S, Venkataraman S. Design and implementation of a robot control system with traded and shared control capability. Proceedings of the IEEE International Conference on Robotics and Automation; 1989. p. 1310–1315.

35. Sellner B, Simmons R, Singh S. User modelling for principled sliding autonomy in human-robot teams. In: Multi-robot systems from swarms to intelligent automata. Berlin (Germany): Springer; 2005;3:197–208.

36. McDowell K, Lin CT, Oie KS, Jung TP, Gordon S, Whitaker KW, Li ST, Lu SW, Hairston WD. Real-world neuroimaging technologies. IEEE Access. 2013;1:131–149.

37. Parasuraman R, Wickens CD. Humans: still vital after all these years of automation. Hum Factors J Hum Factors Ergon Soc. 2008;50(3):511–520.

38. Intraub H. Rapid conceptual identification of sequentially presented pictures. J Exp Psychol Hum Percept Perform. 1981;7(3):604.

39. Mathan S, Ververs P, Dorneich M, Whitlow S, Carciofini J, Erdogmus D, Pavel M, Huang C, Lan T, Adami A. Neurotechnology for image analysis: searching for needles in haystacks efficiently. Augment Cogn Past Present

Future; 2006 [accessed 2014 Sep 5]. http://www3.ece.neu.edu/~erdogmus/publications/C109_AUGCOG2006_NIAsearchingneedles_Santosh.pdf.

40. Parra LC, Christoforou C, Gerson AD, Dyrholm M, Luo A, Wagner M, Philiastides MG, Sajda P. Spatiotemporal linear decoding of brain state. IEEE Signal Process Mag. 2008;25(1):107–15.

41. Files BT, Marathe AR. A regression method for estimating performance in a rapid serial visual presentation target detection task. J Neurosci Methods. 2016;258:114–23.

42. Marathe AR, Ries AJ, Lawhern VJ, Lance BJ, Touryan J, McDowell K, Cecotti H. The effect of target and nontarget similarity on neural classification performance: a boost from confidence. Front Neurosci. 2015;9:270.

43. Sajda P, Pohlmeyer E, Wang J, Parra LC, Christoforou C, Dmochowski J. In a blink of an eye and a switch of a transistor: cortically coupled computer vision. Proc IEEE. 2010;98(3):462–478.

44. Gerson AD, Parra LC, Sajda P. Cortically coupled computer vision for rapid image search. IEEE Trans Neural Syst Rehabil Eng. 2006;14(2):174–179.

45. Marathe AR, Lance BJ, Nothwang W, Metcalfe JS, McDowell K. Confidence metrics improve human-autonomy integration. Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction; 2014.

46. Marathe AR, Ries AJ, McDowell K. A novel method for single-trial classification in the face of temporal variability. In: Schmorrow DD, Fidopiastis CM, editors. Foundations of augmented cognition. Berlin, Heidelberg (Germany): Springer; 2013 p. 345–52.

47. Marathe AR, Ries AJ, McDowell K. Sliding HDCA: single-trial EEG classification to overcome and quantify temporal variability. IEEE Trans Neural Syst Rehabil Eng. 2014;22(2):201–211.

48. Ries AJ, Larkin GB. Stimulus and response-locked P3 activity in a dynamic rapid serial visual presentation (RSVP) task. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2013. Report No.: ARL-TR-6314.

49. US Department of Defense Office, of the Secretary of Defense. Code of federal regulations, protection of human subjects. 32 CFR 219. Washington (DC): Government Printing Office; 1999.

50. US Department of the Army. Use of volunteers as subjects of research. AR 70-25. Washington (DC): Government Printing Office; 1990.

51. Lacouture Y, Cousineau D. How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times. Tutor Quant Methods Psychol. 2008;4(1):35–45.

52. Palmer EM, Horowitz TS, Torralba A, Wolfe JM. What are the shapes of response time distributions in visual search? J Exp Psychol Hum Percept Perform. 2011;37(1):58–71.

53. Raymond JE, Shapiro KL, Arnell KM. Temporary suppression of visual processing in an RSVP task: an attentional blink? J Exp Psychol Hum Percept Perform. 1992;18(3):849–860.

54. Shapiro KL, Raymond JE, Arnell KM. The attentional blink. Trends Cogn Sci. 1997;1(8):291–296.

55. Luo A, Sajda P. Comparing neural correlates of visual target detection in serial visual presentations having different temporal correlations. Front Hum Neurosci; 2009 Apr 21 [accessed 2014 Sep 5]. http://www.ncbi.nlm.nih.gov /pmc/articles/PMC2679198/.

56. Privitera CM, Renninger LW, Carney T, Klein S, Aguilar M. Pupil dilation during visual target detection. J Vis. 2010;10(10):3.

57. Bigdely-Shamlo N, Vankov A, Ramirez RR, Makeig S. Brain activity-based image classification from rapid serial visual presentation. IEEE Trans Neural Syst Rehabil Eng. 2008;16(5):432–441.

58. Sajda P, Gerson A, Parra L. High-throughput image search via single-trial event detection in a rapid serial visual presentation task. Proceedings of the First International IEEE EMBS Conference on Neural Engineering; 2003. p. 7–10.

59. Gerson AD, Parra LC, Sajda P. Cortically coupled computer vision for rapid image search. IEEE Trans Neural Syst Rehabil Eng. 2006;14(2):174–179.

60. Sajda P, Pohlmeyer E, Wang J, Parra LC, Christoforou C, Dmochowski J, Hanna B, Bahlmann C, Singh MK, Chang S. In a blink of an eye and a switch of a transistor: cortically coupled computer vision. Proc IEEE. 2010;98(3):462–478.

61. Cecotti H, Rivet B, Congedo M, Jutten C, Bertrand O, Maby E, Mattout J. A robust sensor-selection method for P300 brain-computer interfaces. J Neural Eng. 2011;8(1):016001.

62. Bigdely-Shamlo N, Vankov A, Ramirez RR, Makeig S. Brain activity-based image classification from rapid serial visual presentation. IEEE Trans Neural Syst Rehabil Eng. 2008;16(5):432–441.

63. Touryan J, Gibson L, Horne JH, Weber P. Real-time measurement of face recognition in rapid serial visual presentation. Front Psychol. 2011;11:2.

64. Touryan J, Gibson L, Horne JH, Weber P. Real-time classification of neural signals corresponding to the detection of targets in video imagery. Procedings of the 3rd International Conference on Applied Human Factors and Ergonomics; 2010. p. 60.

65. Yu K, Shen K, Shao S, Ng WC, Kwok K, Li X. Common spatio-temporal pattern for single-trial detection of event-related potential in rapid serial visual presentation triage. IEEE Trans Biomed Eng. 2011;58(9):2513–2520.

66. Yu K, Shen K, Shao S, Ng WC, Li X. Bilinear common spatial pattern for single-trial ERP-based rapid serial visual presentation triage. J Neural Eng. 2012;9(4):046013.

67. Marathe AR, Ries AJ, McDowell K. A novel method for single-trial classification in the face of temporal variability. In: Schmorrow DD, Fidopiastis CM, editors. Foundations of augmented cognition. Berlin (Germany): Springer; 2013 p. 345–352.

68. Polich J, Comerchero MD. P3a from visual stimuli: typicality, task, and topography. Brain Topogr. 2003;15(3):141–152.

69. Huang Y, Erdogmus D, Mathan S, Pavel M. Large-scale image database triage via EEG evoked responses. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing; 2008. p. 429–432.

70. Mathan S, Erdogmus D, Huang Y, Pavel M, Ververs P, Carciofini J, Dorneich M, Whitlow S. Rapid image analysis using neural signals. In: CHI'08 extended abstracts on human factors in computing systems; 2008. p. 3309–3314 [accessed 2013 Feb 7]. http://dl.acm.org/citation.cfm?id=1358849.

71. Raymond JE, Shapiro KL, Arnell KM. Temporary suppression of visual processing in an RSVP task: an attentional blink? J Exp Psychol Hum Percept Perform. 1992;18(3):849.

72. Chun MM, Potter MC. A two-stage model for multiple target detection in rapid serial visual presentation. J Exp Psychol Hum Percept Perform. 1995;21(1):109–127.

73. Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods. 2004;134(1):9–21.

74. Jung TP, Makeig S, Humphries C, Lee TW, McKeown MJ, Iragui V, Sejnowski TJ. Removing electroencephalographic artifacts by blind source separation. Psychophysiology. 2000;37(2):163–178.

75. Leiva JM, Martens SM. MLSP competition, 2010: description of first place method. Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP); 2010. p. 112–113.

76. Green DM, Swets JA. Signal detection theory and psychophysics. New York (NY): Wiley; 1966.

77. Lopez-Calderon J, Luck SJ. ERPLAB: an open-source toolbox for the analysis of event-related potentials. Front Hum Neurosci. 2014 Apr [accessed 2016 Mar 21]. https://doi.org/10.3389/fnhum.2014.00213.

78. Rivet B, Souloumiac A, Attina V, Gibert G. xDAWN algorithm to enhance evoked potentials: application to brain-computer interface. IEEE Trans on Biomed Eng. 2009;56(8):2035–2043.

79. Hoffmann U, Vesin JM, Ebrahimi T, Diserens K. An efficient P300-based brain-computer interface for disabled subjects. J Neurosci Methods. 2008;167(1):115–125.

80. Cecotti H, Eckstein MP, Giesbrecht B. Effects of performing two visual tasks on single-trial detection of event-related potentials. Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2012. p. 1723–1726.

81. Cecotti H, Marathe A, Ries A. Optimization of single-trial detection of event-related potentials through artificial trials. IEEE Trans on Biomed Eng. 2015;62(9):2170.

82. MacKay DJ. Bayesian interpolation. Neural Comput. 1992;4(3):415–447.

83. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–874.

84. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.

85. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29(4):1165–1188.

86. Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ, Bartlett P, Scholkopf B, Schuurmans D, editors. Advances in large margin classifiers. Cambridge (MA): MIT Press; 2000.

87. Wang J, Pohlmeyer E, Hanna B, Jiang Y-G, Sajda P, Chang SF. Brain state decoding for rapid image retrieval. Proceedings of the 17th ACM International Conference on Multimedia; 2009. p. 945–954.

88. Marathe AR, Lance BJ, Nothwang W, Metcalfe JS, McDowell K. Confidence metrics improve human-autonomy integration. In: Proceedings of the 9th ACM/IEEE international conference on human-robot interaction. New York (NY): IEEE Press; 2014.

89. Steiner GZ, Brennan ML, Gonsalvez CJ, Barry RJ. Comparing P300 modulations: target-to-target interval versus infrequent nontarget-to-nontarget interval in a three-stimulus task. Psychophysiology. 2013;50(2):187–194.

90. Files B, Canady J, Warnell G, Stump E, Nothwang W, Marathe A. Human assisted robotic exploration. Bellingham (WA): SPIE Defense+Security; 2016. p. 98361Y–98361Y.

91. Saeedi S, Trentini M, Seto M, Li H. Multiple-robot simultaneous localization and mapping: a review. J Field Robot. 2016;33(1):3–46.

92. Fuentes-Pacheco J, Ruiz-Ascencio J, Rendón-Mancha JM. Visual simultaneous localization and mapping: a survey. Artif Intell Rev. 2015;43(1):55–81.

93. Cummins M, Newman P. FAB-MAP: probabilistic localization and mapping in the space of appearance. Int J Robot Res. 2008;27(6):647–665.

94. Cummins M, Newman P. Appearance-only SLAM at large scale with FAB-MAP 2.0. Int J Robot Res. 2011;30(9):1100–1123.

95. Cummins M, Newman P. Accelerating FAB-MAP with concentration inequalities. IEEE Trans Robot. 2010;26(6):1042–1050.

96. Se S, Lowe D, Little J. Vision-based mobile robot localization and mapping using scale-invariant features. Proceedings of IEEE International Conference on Robotics and Automation. Piscataway (NJ): IEEE; 2001. p. 2051–2058.

97. Olson E, Strom J, Morton R, Richardson A, Ranganathan P, Goeddel R, Bulic M, Crossman J, Marinier B. Progress toward multi-robot reconnaissance and the MAGIC 2010 competition. J Field Robot. 2012;29(5):762–792.

98. Oliva A, Schyns PG. Diagnostic colors mediate scene recognition. Cognit Psychol. 2000;41(2):176–210.

99. Oliva A, Torralba A. Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis. 2001;42(3):145–175.

100. Greene MR, Oliva A. Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cognit Psychol. 2009;58(2):137–176.

101. Epstein R, Graham KS, Downing PE. Viewpoint-specific scene representations in human parahippocampal cortex. Neuron. 2003;37(5):865–876.

102. Frances Wang R, Simons DJ. Active and passive scene recognition across views. Cognition. 1999;70(2):191–210.

103. Maguire EA, Burgess N, O'Keefe J. Human spatial navigation: cognitive maps, sexual dimorphism, and neural substrates. Curr Opin Neurobiol. 1999;9(2):171–177.

104. Epstein RA, Parker WE, Feiler AM. Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition. J Neurosci. 2007;27(23):6141–6149.

105. Epstein R, DeYoe EA, Press DZ, Rosen AC, Kanwisher N. Neuropsychological evidence for a topographical learning mechanism in parahippocampal cortex. Cogn Neuropsychol. 2001;18(6):481–508.

106. Burgess N. Spatial memory: how egocentric and allocentric combine. Trends Cogn Sci. 2006;10(12):551–557.

107. Shelton AL, Mcnamara TP. Multiple views of spatial memory. Psychon Bull Rev. 1997;4(1):102–106.

108. Wang RF, Spelke ES. Updating egocentric representations in human navigation. Cognition. 2000;77(3):215–250.

109. Wang RF, Spelke ES. Human spatial representation: insights from animals. Trends Cogn Sci. 2002;6(9):376–382.

110. Glover A, Maddern W, Warren M, Reid S, Milford M, Wyeth G. Openfabmap: an open source toolbox for appearance-based loop closure detection. Proceedings of the IEEE International Conference on Robotics and Automation; 2012. p. 4730–4735.

111. Brainard DH. The psychophysics toolbox. Spat Vis. 1997;10(4):433–6.

112. US Department of Defense, Office of the Secretary of Defense. Code of federal regulations, protection of human subjects. 32 CFR 219. Washington (DC): Government Printing Office; 1999.

113. US Department of the Army. Use of volunteers as subjects of research. AR 70-25. Washington (DC): Government Printing Office; 1990.

114. Green DM, Swets JA. Signal detection theory and psychophysics. New York (NY): Wiley; 1966.

115. Mickes L, Wixted JT, Wais PE. A direct test of the unequal-variance signal detection model of recognition memory. Psychon Bull Rev. 2007;14(5):858–865.

116. Milford MJ, Wyeth GF. Mapping a suburb with a single camera using a biologically inspired SLAM system. IEEE Trans Robot. 2008;24(5):1038–53.

117. Maddern W, Milford M, Wyeth G. CAT-SLAM: probabilistic localisation and mapping using a continuous appearance-based trajectory. Int J Robot Res. 2012;31(4):429–451.

118. Lee H, Kwon H, Robinson RM, Nothwang WD, Marathe AM. Dynamic belief fusion for object detection. Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision; 2016 Mar 7–9; Lake Placid, NY. p. 1–9.

119. Robinson RM, Lee H, McCourt MJ, Marathe AR, Kwon H, Ton C, Nothwang WD. Human-autonomy sensor fusion for rapid object detection. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems; 2015. p. 305–312.

120. Lee H, Kwon H, Robinson RM, Donavanik D, Nothwang WD, Marathe AR. Task-conversions for integrating human and machine perception in a unified task. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems; 2016. p. 2751–2758.

121. Fernando B, Fromont E, Muselet D, Sebban M. Discriminative feature fusion for image classification. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2012. p. 3434–3441.

122. Lan X, Ma AJ, Yuen PC. Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 1194–1201.

123. Natarajan P, Wu S, Vitaladevuni S, Zhuang X, Tsakalidis S, Park U, Prasad R, Natarajan P. Multimodal feature fusion for robust event detection in web videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2012. p. 1298–1305.

124. Wang H, Nie F, Huang H, Ding C. Heterogeneous visual features fusion via sparse multimodal machine. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2013. p. 3097–3102.

125. Malisiewicz T, Gupta A, Efros AA. Ensemble of exemplar-svms for object detection and beyond. Proceedings of the IEEE International Conference on Computer Vision; 2011. p. 89–96.

126. Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; p. 886–893.

127. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell. 2010;32(9):1627–1645.

128. Shafer G. A mathematical theory of evidence. Vol. 1. Princeton (NJ): Princeton University Press; 1976.

129. Dempster AP. Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat. 1967;38(2):325–339.

130. Touryan J, Apker G, Lance BJ, Kerick SE, Ries AJ, McDowell K. Estimating Endogenous Changes in Task Performance from EEG. Neuroprosthetics. 2014;8:155.

131. Everingham M, Van Gool L, Williams CK, Zisserman A. The PASCAL visual object classes challenge 2007 (VOC2007); 2007 [accessed 2017 Mar 21]. http://www.pascal-network.org/challenges/VOC/voc2007/index.html.

132. Kwon J, Lee KM. Visual tracking decomposition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2010. p. 1269–1276.

133. Kwon J, Lee KM. Tracking by sampling trackers. Proceedings of the IEEE International Conference on Computer Vision; 2011. p. 1195–1202.

134. Wu S, Bondugula S, Luisier F, Zhuang X, Natarajan P. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 2665–2672.

135. Bailer C, Pagani A, Stricker D. A superior tracking approach: building a strong tracker through fusion. In: Proceedings of the European Conference on Computer Vision. Berlin (Germany): Springer; 2014. p. 170–185.

136. Kim T, Lee H, Lee K. Optical flow via locally adaptive fusion of complementary data costs. Proceedings of the IEEE International Conference on Computer Vision; 2013. p. 3344–3351.

137. Liu D, Lai KT, Ye G, Chen MS, Chang SF. Sample-specific late fusion for visual category recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2013. p. 803–810.

138. Ma AJ, Yuen PC. Reduced analytical dependency modeling for classifier fusion. In: Proceedings of the European Conference on Computer Vision. Berlin (Germany): Springer; 2012. p. 792–805.

139. Liu C, Yuen J, Torralba A. Sift flow: dense correspondence across scenes and its applications. IEEE Trans on Pattern Anal Mach Intell. 2011;33(5):978–994.

140. Khan FS, Anwer RM, Van De Weijer J, Bagdanov AD, Vanrell M, Lopez AM. Color attributes for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2012. p. 3306–3313.

141. Wang J, Jebara T, Chang SF. Graph transduction via alternating minimization. Proceedings of the 25th International Conference on Machine Learning; 2008. p. 1144–1151.

142. Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 1717–1724.

143. Girshick R. Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 1440–1448.

144. Liu J, McCloskey S, Liu Y. Local expert forest of score fusion for video event classification. In: European conference on computer vision. Berlin (Germany): Springer; 2012. p. 397–410.

145. Karaoglu S, Liu Y, Gevers T. Detect2rank: combining object detectors using learning to rank. IEEE Trans Image Process. 2016;25(1):233–248.

146. Xu L, Krzyzak A, Suen CY. Methods of combining multiple classifiers and their applications to handwriting recognition. IEEE Trans Syst Man Cybern. 1992;22(3):418–435.

147. Hoiem D, Chodpathumwan Y, Dai Q. Diagnosing error in object detectors. In: European onference on computer vision. Berlin (Germany): Springer; 2012. p. 340–353.

148. Everingham M, Eslami SA, Van Gool L, Williams CK, Winn J, Zisserman A. The Pascal visual object classes challenge: a retrospective. Int J Comput Vis. 2015;111(1):98–136.

149. Ahonen T, Hadid A, Pietikäinen M. Face recognition with local binary patterns. In: European conference on computer vision. Berlin (Germany): Springer; 2004. p. 469–481.

150. Cevikalp H, Triggs B. Efficient object detection using cascades of nearest convex model classifiers. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2012. p. 3138–3145.

151. Cevikalp H, Triggs B, Franc V. Face and landmark detection by using cascade of classifiers. Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition; 2013. p. 1–7.

152. Yao B, Jiang X, Khosla A, Lin AL, Guibas L, Fei-Fei L. Human action recognition by learning bases of action attributes and parts. Proceedings of the IEEE International Conference on Computer Vision; 2011. p. 1331–1338.

153. Fanaee-T H, Gama J. Event labeling combining ensemble detectors and background knowledge. Prog Artif Intell. 2014;2(2–3):113–127.

154. Millán JdR, Rupp R, Müller-Putz GR, Murray-Smith R, Giugliemma C, Tangermann M, Vidaurre C, Cincotti F, Kübler A, Leeb R, et al. Combining brain-computer interfaces and assistive technologies: state-of-the-art and challenges. Front Neurosci. 2010;4:161.

155. White JR, Levy T, Bishop W, Beaty JD. Real-time decision fusion for multimodal neural prosthetic devices. PloS One. 2010;5(3):e9493.

156. Huang H, Zhang F, Hargrove LJ, Dou Z, Rogers DR, Englehart KB. Continuous locomotion-mode identification for prosthetic legs based on neuromuscular–mechanical fusion. IEEE Trans Biomed Eng. 2011;58(10):2867–2875.

157. Sajda P, Gerson A, Parra L. High-throughput image search via single-trial event detection in a rapid serial visual presentation task; 2003. p. 7–10 [accessed 2013 Jan 31]. http://ieeexplore.ieee.org/xpls/abs_all.jsp ?arnumber=1196297.

158. Ramoser H, Muller-Gerking J, Pfurtscheller G. Optimal spatial filtering of single trial EEG during imagined hand movement. IEEE Trans Rehabil Eng. 2000;8(4):441–6.

159. Potter MC, Levy EI. Recognition memory for a rapid sequence of pictures. J Exp Psychol. 1969;81(1):10.

160. Robeck MC, Wallace RR. The psychology of reading: an interdisciplinary approach. Mahwah (NJ): Lawrence Erlbaum; 1990.

161. Kruse AA. Neurotechnology for intelligence analysts. In: Gardner PJ, Fountain AW 3rd, editors. Chemical and biological sensing VII: Proceedings of the SPIE. 2006;2618 [accessed 2016 Feb 16]. doi: 10.1117/12.666054.

162. Mathan S, Ververs P, Dorneich M, Whitlow S, Carciofini J, Erdogmus D, Pavel M, Huang C, Lan T, Adami. Neurotechnology for image analysis: searching for needles in haystacks efficiently. In: Augmented cognition: past, present, and future. Arlington (VA): Strategic Analysis, Inc.; 2006.

163. Fei-Fei L, Iyer A, Koch C, Perona P. What do we perceive in a glance of a real-world scene? J Vis. 2007;7(1):10.

164. Evans KK, Treisman A. Perception of objects in natural scenes: is it really attention free? J Exp Psychol Hum Percept Perform. 2005;31(6):1476.

165. Branson S, Wah C, Schroff F, Babenko B, Welinder P, Perona P, Belongie S. Visual recognition with humans in the loop. In: Proceedings of the European Conference on Computer Vision. Berlin (Germany): Springer; 2010. p. 438–451.

166. Kapoor A, Grauman K, Urtasun R, Darrell T. Active learning with Gaussian processes for object categorization. Proceedings of the IEEE 11th International Conference on Computer Vision; 2007. p. 1–8.

167. Holub A, Perona P, Burl MC. Entropy-based active learning for object recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops; 2008. p. 1–8.

168. Marathe AR, Lawhern VJ, Wu D, Slayback D, Lance BJ. Improved neural signal classification in a rapid serial visual presentation task using active learning. IEEE Trans Neural Syst Rehabil Eng. 2016;24(3):333–343.

169. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Proceedings of the Conference on Neural Information Processing Systems; 2012. p. 1097–1105.

170. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis. 2015;115(3):211–252.

171. Mathis A, Nothwang W, Donavanik D, Conroy J, Shamwell J, and Robinson R. Making optic flow robust to dynamic lighting conditions for real time operation. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2016 Mar. Report No.: ARL-TR-7629.

172. Micire M. Fast lightweight autonomy (FLA). Arlington (VA): Defense Science Office, Defense Advanced Research Projects Agency; 2014 Dec 22.

173. Bohren J, Foote T, Keller J, Kushleyev A, Lee D, Stewart A, Vernaza P, Derenick J, Spletzer J, Satterfield B. Little Ben: the Ben Franklin racing team's entry in the 2007 DARPA urban challenge. J Field Robot. 2008;25(9):598–614.

174. Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. Proceedings of the International Joint Conference on Artificial Intelligence; 1981. p. 674–679.

175. Krapp HG, Hengstenberg R. Estimation of self-motion by optic flow processing in single visual interneurons. Nature. 1996;384(6608):463–466.

176. Horn BK, Schunck BG. Determining optical flow. In: Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series; 1981. p. 319–331.

177. Schlyter P. How bright are natural light sources; 2006 [accessed 2016 Mar 21]. http://www.stjarnhimlen.se/comp/radfaq.html.

178. Dederscheck D, Müller T, Mester R. Illumination invariance for driving scene optical flow using comparagram preselection. Proceedings of the IEEE Intelligent Vehicles Symposium; 2012. p. 742–747.

179. Zhang L, Sakurai T, Miike H. Detection of motion fields under spatio-temporal non-uniform illumination. Image Vis Comput. 1999;17(3):309–320.

180. Mileva YM. Invariance with optic flow. Saarbrücken (Germany): Department of Computer Science, Saarland University; 2007.

181. Mileva Y, Bruhn A, Weickert J. Illumination-robust variational optical flow with photometric invariants. In: Joint pattern recognition symposium. Berlin (Germany): Springer; 2007. p. 152–162.

182. Van De Weijer J, Gevers T, Smeulders AW. Robust photometric invariant features from the color tensor. IEEE Trans Image Process. 2006;15(1):118–127.

183. Schuchert T, Aach T, Scharr H. Range flow in varying illumination: algorithms and comparisons. IEEE Trans Pattern Anal Mach Intell. 2010;32(9):1646–1658.

184. Kendoul F, Fantoni I, Nonami K. Optic flow-based vision system for autonomous 3D localization and control of small aerial vehicles. Robot Auton Syst. 2009;57(6):591–602.

185. Haussecker HW, Fleet DJ. Computing optical flow with physical models of brightness variation. IEEE Trans Pattern Anal Mach Intell. 2001;23(6):661–673.

186. Zimmer H, Bruhn A, Weickert J, Valgaerts L, Salgado A, Rosenhahn B, Seidel H-P. Complementary optic flow. In: International workshop on energy minimization methods in computer vision and pattern recognition. Berlin (Germany): Springer; 2009. p. 207–220.

187. Kearney JK, Thompson WB, Boley DL. Optical flow estimation: an error analysis of gradient-based methods with local optimization. IEEE Trans Pattern Anal Mach Intell. 1987;9(2):229–244.

188. Enkelmann W. Obstacle detection by evaluation of optical flow fields from image sequences. Image Vis Comput. 1991;9(3):160–168.

189. Christmas WJ. Filtering requirements for gradient-based optical flow measurement. IEEE Trans Image Process. 2000;9(10):1817–1820.

190. Lempitsky V, Roth S, Rother C. Fusionflow: discrete-continuous optimization for optical flow estimation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2008. p. 1–8.

191. Sellent A, Eisemann M, Goldlücke B, Pock T, Cremers D, Magnor MA. Variational optical flow from alternate exposure images. Proceedings of the Vision, Modeling and Visualization Conference; 2009. p. 135–144.

192. Sharmin N, Brad R. Optimal filter estimation for Lucas-Kanade optical flow. Sensors. 2012;12(9):12694–12709.

193. Baker S, Matthews I. Lucas-Kanade 20 years on: a unifying framework. Int J Comput Vis. 2004;56(3):221–255.

194. Kameda Y, Imiya A, Ohnishi N. A convergence proof for the Horn-Schunck optical-flow computation scheme using neighborhood decomposition. In: International Workshop on Combinatorial Image Analysis. Berlin (Germany): Springer; 2008. p. 262–273.

195. Anderson M, Iandola F, Keutzer K. Quantifying the energy efficiency of object recognition and optical flow. Berkeley (CA): University of California at Berkeley; 2014. Technical Report No.: UCB/EECS-2014-22.

196. Bruhn A, Weickert J, Schnörr C. Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. Int J Comput Vis. 2005;61(3):211–231.

197. Bilgic B. Iterative pyramidal LK optical flow. MathWorks; 2009 [accessed 2016 Feb 16]. https://www.mathworks.com/matlabcentral/fileexchange/23142-iterative-pyramidal-lk-optical-flow?requestedDomain=www.mathworks.com.

198. Kharbat M, Horn-Schunck. Optical flow method; 2009. [accessed 2016 Feb 16]. https://wwwmathworks.com/matlabcentral/fileexchange/22756-horn-schunck-optical-floq-method.

199. Brayboy J. Research lab to test insect-inspired robotic platform in Army exercise. ARL News; 2014 Nov 7.

209. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2012. p. 3354–3361.

201. Butler DJ, Wulff J, Stanley GB, Black MJ. A naturalistic open source movie for optical flow evaluation. In: European conference on computer vision. Berlin (Germany): Springer; 2012. p. 611–625.

202. Ryan D, Denman S, Fookes C, Sridharan S. Textures of optical flow for real-time anomaly detection in crowds. Proceedings of the IEEE International Conference Advanced Video and Signal-Based Surveillance; 2011. p. 230–235.

203. Mehran R, Oyama A, Shah M. Abnormal crowd behavior detection using social force model. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2009. p. 935–942.

204. Conroy J, Gremillion G, Ranganathan B, Humbert JS. Implementation of wide-field integration of optic flow for autonomous quadrotor navigation. Auton Robots. 2009;27(3):189–198.

205. Hyslop AM, Humbert JS. Autonomous navigation in three-dimensional urban environments using wide-field integration of optic flow. J Guid Control Dyn. 2010;33(1):147–159.

206. Bideau P, Learned-Miller E. It's moving! A probabilistic model for causal motion segmentation in moving camera videos. In: European Conference on Computer Vision. Berlin (Germany): Springer; 2016. p. 433–449.

207. Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: the KITTI dataset. Int J Robot Res. 2013;32(11):1231–1237.

208. Martin D, Fowlkes C, Tal D, Malik J. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Proceedings of the IEEE International Conference on Computer Vision; 2001. p. 416–423.

209. Weiss S, Achtelik MW, Lynen S, Chli M, Siegwart R. Real-time onboard visual-inertial state estimation and self-calibration of mavs in unknown environments. Proceedings of the IEEE International Conference on Robotics and Automation; 2012. p. 957–964.

210. Shen S, Mulgaonkar Y, Michael N, Kumar V. Vision-based state estimation for autonomous rotorcraft mavs in complex environments. Proceedings of the IEEE International Conference on Robotics and Automation; 2013. p. 1758–1764.

211. Klein G, Murray D. Parallel tracking and mapping for small AR workspaces. Proceedings of the IEEE and ACM International Symposium on Mixed and Augmented Reality; 2007. p. 225–234.

212. Koppel A, Warnell G, Stump E, Ribeiro A. D4l: decentralized dynamic discriminative dictionary learning. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems; 2015. p. 2966–2973.

213. Koppel A, Warnell G, Stump E. Task-driven dictionary learning in distributed online settings. Proceedings of the 49th IEEE Asilomar Conference on Signals, Systems and Computers; 2015. p. 1114–1118.

214. Koppel A, Warnell G, Stump E, Ribeiro A. D4L: decentralized dynamic discriminative dictionary learning. IEEE Trans Signal Inf Process Netw Prep.; 2016.

215. Koppel A, Warnell G, Stump E, Ribeiro A. Parsimonious online kernel learning via sparse projections in function space. J Mach Learn Res Rev.; 2016 submitted.

216. Koppel A, Warnell G, Stump E, Ribeiro A. Parsimonious Online Kernel Learning via Sparse Projections in Function Space. IEEE Conf Audio Speech Signal Process Rev.; 2016 submitted.

217. Koppel A, Fink J, Warnell G, Stump E, Ribeiro A. Online learning for characterizing unknown environments in ground robotic vehicle models. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems; 2016. p. 626–633.

218. Skodras A, Christopoulos C, Ebrahimi T. The JPEG 2000 still image compression standard. IEEE Signal Process Mag. 2001;18(5):36–58.

219. Files BT, Marathe AR. A regression method for estimating performance in a rapid serial visual presentation target detection task. J Neurosci Methods. 2016;258:114–23.

220. Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods. 2004;134(1):9–21.

221. Bell AJ, Sejnowski TJ. An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 1995;7(6):1129–1159.

222. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell. 2010;32(9):1627–1645.

223. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 580–587.

List of Symbols, Abbreviations, and Acronyms

ANOVA	analysis of variance
AP	average precision
ARL	US Army Research Laboratory
AUC	area under the curve
Az	receiver operating characteristic curve
B	background distractor
bel(*A*)	belief function
BCI	brain–computer interface
BLDA	Bayesian linear discriminant analysis
BG	background
BP	button press
BPA	basic probability assignment
CNN	convolutional neural network
CSP	common spatial patterns
CV	computer vision
d´	d-prime
D2R	Detect2Rank
D4L	Decentralized Dynamic Discriminative Dictionary Learning
DARPA	Defense Advanced Research Projects Agency
DBF	Dynamic Belief Fusion
DOD	US Department of Defense
DPM	deformable part model
DSI	Director's Strategic Initiative
DSIFT	dense scale-invariant feature transform
DST	Dempster–Shafer Theory

EEG	electroencephalography
EO	electro-optical
EOG	electro-oculogram
ERP	event-related potential
ESVM	exemplar support vector machine
FAR	false alarm rate
FP	false positive
FTCNN	fine-tuned convolutional neural network
FY	fiscal year
GPS	global positioning system
HAI	human–autonomy interaction
HDCA	hierarchical discriminant component analysis
HOG	histogram of oriented gradient
HR	hit rate
HSV	hue saturation value
ICA	independent component analysis
IMU	inertial measurement unit
LED	light-emitting diode
LEF	local expert forest
LiIDAR	light detection and ranging
Loc	localization
lux	one lumen per square meter
mAP	mean average precision
MCOE	US Army Maneuver Center for Excellence
MNIST	Mixed National Institute of Standards and Technology
MOUT	military operations in urban terrain
NC	neural classifier

NT	nontarget
OLDA	Online Learning for Drivability Assessment
Oth	dissimilar
PNT	Position, Navigation, and Timing
POLK	Parsimonious Online Learning with Kernels
PR	precision recall
RCNN	regional convolutional neural network
RGB	red, green, blue
ROC	receiver operating characteristic
RSVP	rapid serial visual presentation
RT-PDF	response-time probability density function
SDT	Signal Detection Theory
SIFT	scale-invariant feature transform
Sim	similar
SSNR	signal to signal plus noise ratio
SVM	support vector machine
SWaP	size, weight, and power
T	target
TAG	transductive annotation by graph
TN	Target and Non Target
TO	Target Only
TOC	tactical operations center
UAS	unmanned aerial system
UAV	unmanned aerial vehicle
UGV	unmanned ground vehicle
WS	weighted sum
XD	xDawn/XD+BLDA

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA

2 (PDF) DIRECTOR US ARMY RESEARCH LAB RDRL CIO L IMAL HRA MAIL & RECORDS MGMT

1 (PDF) GOVT PRINTG OFC A MALHOTRA

1 (PDF) DIRECTOR US ARMY RESEARCH LAB RDRL HRF D A MARATHE